Check for updates

The extended persistent homology transform of manifolds with boundary

Katharine Turner[1] · Vanessa Robins[1] · James Morgan[2]

Abstract

A shortcoming of persistent homology is that when two domains have different numbers of components or holes the persistence diagrams of any filtration will have an infinite distance between them. We address this issue by revisiting the theory of extended persistence, initially developed by Cohen-Steiner, Edelsbrunner and Harer in 2009 to quantify the support of the essential homology classes for Morse functions on manifolds. We simplify the mathematical treatment of extended persistence by formulating it as a persistence module derived from a sequence of relative homology groups for pairs of spaces. Then, for n-manifolds with boundary embedded in $\mathbb{R}^n$, we use Morse theory to show that the extended persistent homology of a height function over M can be deduced from the extended persistent homology of the same height function restricted to ∂M. As an application, we describe the extended persistent homology transform (XPHT); a topological transform which takes as input a shape embedded in Euclidean space, and to each unit vector assigns the extended persistence module of the height function over that shape with respect to that direction. We define a distance between two shapes by integrating over the sphere the distance between the respective extended persistence modules. By using extended persistence we get finite distances between shapes even when they have different numbers of essential classes. We study the application of the XPHT to binary images; outlining an algorithm for efficient calculation of the XPHT exploiting relationships between the PHT of the boundary curves to the extended persistence of the foreground.

✉ Katharine Turner
katharine.turner@anu.edu.au

1 Australian National University, Canberra, ACT, Australia

2 University of Sydney, Sydney, NSW, Australia

🖄 Springer

1 Introduction

The fundamental goal in statistical shape analysis is to define and compute meaningful distances between different subsets of Euclidean space. A recent landmark-free approach to quantify both the geometry and topology of a shape is to use a topological transform such as the Persistent Homology Transform (PHT) or the Euler Characteristic Transform (ECT). Both of these transforms take a shape M, viewed as a subset $\mathbb{R}^n$, and associate to each direction $v \in S^{n-1}$ a shape summary obtained by scanning M in the direction v, calculating the persistent homology ($\mathrm{PH}(M, v)$) and the Euler curve respectively.

Different formulations of the PHT and ECT have been demonstrably useful in diverse applications including prediction of disease progression from the shapes of tumours (Crawford et al. 2020; Shboul et al. 2019), identification of different cultivars from the shapes of leaves (Zhang et al. 2021), quantification of morphological variation of barley seeds (Amézquita et al. 2022), and identification of structural differences among proteins (Tang et al. 2022). This paper introduces an improved variant of this topological transform called the extended persistent homology transform (XPHT) and establishes properties that significantly reduce the time required to compute it.

A limitation of the PHT is it does not work well with shapes that have different Betti numbers (the ranks of the homology groups). For $M_1, M_2 \subset \mathbb{R}^n$, the p-distance between their persistent homology transforms is defined as

$$\mathrm{dist}_p(\mathrm{PHT}(M_1), \mathrm{PHT}(M_2))^p = \int_{S^{n-1}} W_p(\mathrm{PH}(M_1, v), \mathrm{PH}(M_2, v))^p \, dv$$

where $W_p(\cdot, \cdot)$ is the p-Wasserstein distance. If M_1 and M_2 have different Betti numbers, then

$$W_p(\mathrm{PH}(M_1, v), \mathrm{PH}(M_2, v)) = \infty,$$

for all v, and thus

$$\mathrm{dist}_p(\mathrm{PHT}(M_1), \mathrm{PHT}(M_2)) = \infty.$$

One potential work-around would be to replace the Wasserstein distance with a different metric on the space of persistence modules, one where having different Betti numbers does not enforce infinite distance. A more satisfying approach is to replace persistent homology with extended persistent homology.

The theory of extended persistence for functions over a manifold X was developed in Cohen-Steiner et al. (2009b) to quantify the support of the essential homology classes of X (these essential classes are the elements of $H_*(X)$). Even when the domains have different Betti numbers we still have a finite Wasserstein distance between their extended persistence modules. This motivates the Extended Persistent Homology Transform (XPHT) as a topological transform, which is defined in exactly the same manner as the PHT but replacing regular persistent homology with extended persistent homology. By quantifying the size of essential classes it is possible for the XPHT to be stable with respect to the addition or removal of "small" essential classes in the different domains. For example, if we add an isolated noisy pixel to a binary

image then the change in the XPHT will be commensurate with the size of a pixel. This extra stability can provide greater power and robustness to statistical methods that use distances between shapes derived from the XPHT. As this paper is focused on computational aspects of the XPHT, comprehensive stability results are left as a future research direction.

We believe that extended persistence is currently under-utilised within applied topology and this paper addresses three potential obstacles. Firstly, we make extended persistence modules more theoretically accessible by placing them within a generalised framework that includes both regular persistence as well as extended persistence. Secondly, we provide motivation with an important example (in the form of the XPHT) where using extended persistence provides a qualitative improvement in usefulness. Lastly, we provide insights on how to ease the computation of extended persistence in the important case of height functions, with implemented code for binary images.

1.1 Outline of paper

The mathematical treatment of the XPHT and algorithms to compute it requires the adaption and extension of many standard definitions within applied topology. We cover this material in some detail to make the paper more self-contained and to provide a cohesive perspective on results from different areas of the literature.

The original definition of extended persistence in Cohen-Steiner et al. (2009b) is made for functions defined on a smooth or piecewise-linear (PL) manifold and concatenates two homology sequences, the standard inclusion-induced persistent homology sequence for the sublevel set filtration, followed by a descending relative homology sequence for superlevel sets. In Sect. 2, we reformulate this as a persistence module over a totally ordered set, with all transition maps defined as those induced on relative homology by inclusions of a pair of spaces. These spaces are defined by a real-valued function on a triangulated manifold with boundary, $f : M \to \mathbb{R}$. We then establish a relationship between the intervals of extended persistence modules of f and $(-f)$, which is one of the results required to reduce computation time for the XPHT.

In Sect. 3 we generalise the definition of Wasserstein and bottleneck distances between persistence diagrams to apply to persistence modules over a totally ordered metric space, with a defined set of ephemeral (zero-length) intervals. The Wasserstein and bottleneck distances are optimal transport metrics with transport plans that include a bijection between chosen subsets of intervals and then subsets of unmatched intervals. To define the cost of a transportation plan we need a distance between intervals and cost of having an interval unmatched. We show our definition agrees with the existing definitions of bottleneck distance between extended persistence diagrams.

A key theoretical insight of our work, and one which makes the XPHT feasible to compute, is that for manifolds with boundary embedded in $\mathbb{R}^n$ the extended persistent homology of a height function over M can be deduced from the extended persistent homology of the same height function restricted to ∂M. This is the topic of Sect. 4. The proof of this insight requires ideas from Morse theory for manifolds with boundary, in both the smooth and piecewise-linear settings. This background material is covered in Sect. 4.1. We also precisely state the relationship between birth and death parameters of extended persistence in terms of the different kinds of critical points of a smooth

or PL Morse function on a manifold with boundary. Section 4.2 then develops results specifically for the case of a directional height function. It is worth noting that any subset of $\mathbb{R}^n$ with positive weak feature size is arbitrarily close to a n-manifold with boundary by taking an ϵ expansion. This means the restriction to n-manifolds with boundary is reasonable from an application standpoint.

Adapting the definition of the persistent homology transform (PHT) to extended persistence is straightforward. We cover this material in Sect. 5.

Shape analysis of objects in digital images is an application domain with wide interest. Objects in binary images can be modelled as two dimensional manifolds with boundary lying in the plane, so our XPHT results apply. In Sect. 6 we define boundary curves that separate foreground and background connected components consistent with a chosen digital adjacency, and show that these boundary curves are disjoint simple closed PL 1-manifolds. Digital grids create degeneracy in the height function critical values, so we derive additional results that establish the correctness of our implemented algorithms. Finally, in Sect. 7 we illustrate our R-package implementation by comparing the XPHT of the letters 'A' and 'g' rendered in a variety of standard fonts. We find the XPHT of the upper case 'A' naturally separates the serif and sans-serif fonts, and that the XPHT of the lower case 'g' naturally separates the single-storey and the double-storey fonts.

1.2 Relation to Alexander duality for extended persistence

A form of Alexander Duality for extended persistence was proved in Edelsbrunner and Kerber (2012). That paper considers the decomposition of the sphere into two sets U, V with $U \cup V = S^n$ and $U \cap V$ a $(n-1)$-manifold, and proves results about the extended persistence of a perfect Morse function f over these sets. A perfect Morse function over S^n is a smooth function with exactly two critical points, one minimum and one maximum. Edelsbrunner and Kerber prove that the extended persistence module of $U \cap V$ is the direct sum of those for U and V (with minor adjustments for degree-0 homology). The statement of our Theorem 4.17 is effectively a special case of their result. However, our proof is very different as it is based on Morse theory instead of Alexander Duality. Another key difference in our results is that we show how the extended persistence module for $U \cap V$ splits into the two different parts (Theorem 4.18); this is not established in Edelsbrunner and Kerber (2012). Since our ultimate goal is to calculate the extended persistence of U from that of $U \cap V$ this splitting criterion is pivotal.

2 Extended persistence modules

2.1 Persistence modules over totally ordered sets

Commonly, persistence modules are defined with an underlying parameter space a subset of $\mathbb{R}$ but they can be defined where the parameter space is any totally ordered set. This approach makes working with extended persistence substantially cleaner and more intuitive as it facilitates the split of the single parameter space into ordinary and relative homology parameter types.

Definition 2.1 A *totally ordered set* $(\Theta, \leq)$ is a set Θ with a relation $\leq$ which is

- Reflexive: that is $\alpha \leq \alpha$ for all $\alpha \in \Theta$,
- Antisymmetric: that is $\alpha \leq \beta$ and $\beta \leq \alpha$ implies $\alpha = \beta$,
- Transitive: that is $\alpha \leq \beta$ and $\beta \leq \gamma$ implies $\alpha \leq \gamma$, and
- Comparable: for all α, β either $\alpha \leq \beta$ or $\beta \leq \alpha$.

Definition 2.2 Fix a field $\mathbb{F}$ and let Θ be a totally ordered set. A *persistence module* $\mathcal{P}$ over Θ is a family $\{V_\alpha\}_{\alpha \in \Theta}$ of $\mathbb{F}$-vector spaces indexed by elements of Θ, together with a family of homomorphisms $\{\varphi_\alpha^\beta : V_\alpha \to V_\beta\}$ such that $\varphi_\alpha^\gamma = \varphi_\beta^\gamma \circ \varphi_\alpha^\beta$ for all $\alpha \leq \beta \leq \gamma$, and $\varphi_\alpha^\alpha = \mathrm{id}_{V_\alpha}$. We call the φ_α^β *transition maps*. We say $\mathcal{P}$ is *pointwise finite dimensional* if the V_α are finite dimensional for all $\alpha \in \Theta$.

In the algebraic theory of persistence modules there are often technical requirements about tameness, and being pointwise finite dimensional is generally a sufficient condition. This is a reasonable assumption in almost any application. The most important algebraic result is the decomposition theorem. This gives a complete yet discrete description of a persistence module up to isomorphism. We will decompose persistence modules into sums of interval modules, but first we must define interval persistence modules.

We are all familiar with intervals that are subsets of the real line. We generalise this notion to any totally ordered set as follows.

Definition 2.3 An *interval* in a totally ordered space $(\Theta, \leq)$ is a subset $I \subset \Theta$ such that for all $\alpha \in \Theta$ either $\alpha \in I$, or $\alpha \leq \theta$ for all $\theta \in I$, or $\theta \leq \alpha$ for all $\theta \in I$. An *interval module* over an interval I is a persistence module $\mathcal{I}_I$ with vector spaces

$$
V_\alpha = \begin{cases} \mathbb{F} \text{ for } \alpha \in I \\ 0 \text{ for } \alpha \notin I \end{cases}
$$

and transition maps $\varphi_\alpha^\beta = \mathrm{id}_\mathbb{F}$ when both $\alpha, \beta \in I$ and 0 otherwise.

For each interval module $\mathcal{I}_I$ we call $\mathfrak{b}(\mathcal{I}_I) = \inf I$ the *birth* parameter and $\mathfrak{d}(\mathcal{I}_I) = \sup I$ the *death* parameter.

The nomenclature of "interval" was introduced for persistence modules with parameter space $\mathbb{R}$ but it is still reasonable even in the generalised setting of totally ordered sets. If we can map the totally ordered set to a subset of the real line, say $f : \Theta \to \mathbb{R}$, in a way that respects the order relation, then we can view each interval module as having support $f^{-1}(I)$ where $I \subset \mathbb{R}$ is some interval.

Theorem 2.4 (Crawley-Boevey (2015) Theorem 1.1) *A pointwise finite dimensional persistence module over any subset of $\mathbb{R}$ admits an* interval decomposition*. That is, there is a multiset of intervals S such that the module is isomorphic to a direct sum of interval modules*

$$
\bigoplus_{I \in S} \mathcal{I}_I
$$

where each $\mathcal{I}_I$ is an interval module. This decomposition is unique up to isomorphism.

For the rest of the paper we assume all persistence modules are pointwise finite dimensional and that the underlying parameter space is equivalent to a subset of $\mathbb{R}$ (with respect to the order relation), and thus we can always assume an interval decomposition occurs.

Given a persistence module $\mathcal{P} = \bigoplus_{I \in S^{\mathcal{P}}} \mathcal{I}_I$ we will use

$$\mathfrak{b}(\mathcal{P}) = \{\mathfrak{b}(\mathcal{I}_I) : I \in S^{\mathcal{P}}\} \quad \text{and} \quad \mathfrak{d}(\mathcal{P}) = \{\mathfrak{d}(\mathcal{I}_I) : I \in S^{\mathcal{P}}\}$$

to denote the multiset of birth parameters and death parameters in the interval decomposition of $\mathcal{P}$.

Readers may be familiar with the terms persistence barcode and persistence diagram. Barcodes and diagrams are graphical representations of the interval decomposition of a persistence module. In particular, a *persistence diagram* consists of a multiset of points in $\mathbb{R}^2$, with each point (x, y) recording the birth and death parameters of an interval from the decomposition. We use all three terms in this paper.

2.2 Extended persistence

Extended persistence combines the regular filtration of sublevel sets for $f : M \to \mathbb{R}$ with a filtration of *relative* homology groups of M with respect to superlevel sets of f. This provides a wealth of extra information about the structure of M, especially in the case that M is a manifold with boundary.

We first recall the definition of relative homology, and the maps induced by the inclusion of a pair. Given a subcomplex $X \subset Y$ we observe that the boundary map on $C_*(Y)$ leaves $C_*(X)$ invariant. This means we can define a chain complex $C_*(Y, X)$ where $C_k(Y, X) = C_k(Y)/C_k(X)$ and the boundary map is

$$\partial_k^{(Y,X)}(\alpha + C_k(X)) = \partial(\alpha) + C_{k-1}(X).$$

We can then define the relative homology groups by

$$H_k(Y, X) = \ker \partial_k^{(Y,X)} / \operatorname{im} \partial_{k+1}^{(Y,X)}.$$

Relative homology is a generalisation of normal homology as $H_k(Y) = H_k(Y, \emptyset)$.

If $X \subset Y \subset B$ and $X \subset A \subset B$ we have an inclusion of pairs $(Y, X) \subset (B, A)$. This inclusion of pairs induces a map between their relative homology groups, $\iota_* : H_k(Y, X) \to H_k(B, A)$, with $\iota(\alpha + C_k(X)) = \alpha + C_k(A)$.

We are now ready to define the extended persistent homology module as a form of persistence module. The parameter space over which this module is constructed is the union of two sets—one corresponding to ordinary homology and the other corresponding to relative homology. Set $O = \{(t, \mathrm{Ord}) : t \in \mathbb{R}\}$ and $R = \{(t, \mathrm{Rel}) : t \in \mathbb{R}\}$. Let $\Theta = O \cup R$. We define a total order over Θ by

$$(s, \mathrm{Ord}) < (t, \mathrm{Ord}) \text{ when } s < t$$
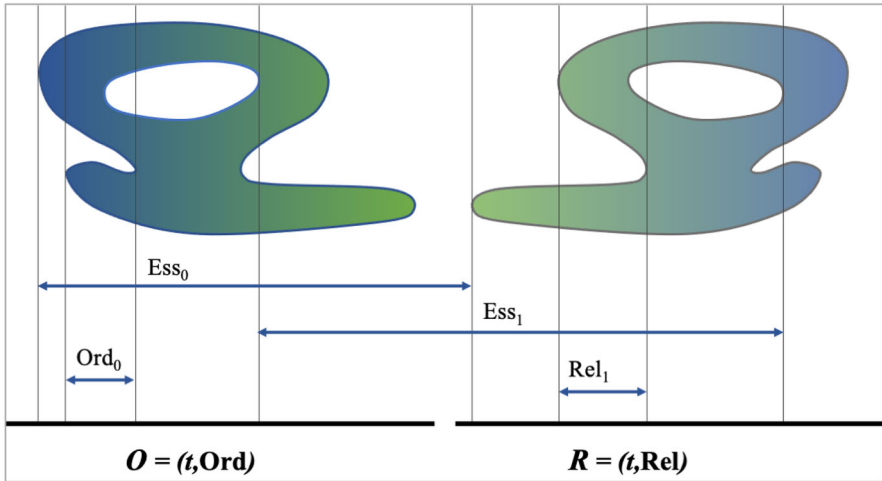
Fig. 1 An illustration of extended persistence intervals for a rather abstract snail, M. The function $f : M \to \mathbb{R}$ is simply the x-coordinate and the function value is denoted by the blue-green colour gradient. We have drawn a copy of M with its x-coordinate reflected to illustrate the superlevel sets used in the relative part of the sequence (color figure online)

$$(s, \text{Rel}) < (t, \text{Rel}) \text{ when } s > t$$
$$(s, \text{Ord}) < (t, \text{Rel}) \text{ for all } s, t$$

We then assign vector spaces to each $\theta \in \Theta$ defined by taking homology of suitable pairs of sublevel and superlevel sets. As input we have a topological space M with a bounded function $f : M \to \mathbb{R}$. Let $M_s = f^{-1}(-\infty, s]$ and $M^s = f^{-1}[s, \infty)$ denote the sublevel and superlevel sets of $f : M \to \mathbb{R}$. We assign the vector spaces as $V_{(t, \text{Ord})} = H_k(M_t, \emptyset)$ and $V_{(t, \text{Rel})} = H_k(M, M^t)$. The transition maps are the natural ones induced by inclusions of a pair. The composition of two such transition maps corresponds exactly to the map on homology induced by the composition of inclusions. This means that the transition maps commute as needed and we have constructed a persistence module.

Each element of the interval decomposition will be supported over some interval of Θ. These intervals are of three types. If the support contains only parameters in O we call it *ordinary*, if the support is a subset of parameters in R we call it *relative*. Finally, the persistent homology class might exist for parameters spanning both O and R, in which case we call it *essential*. Essential persistent homology classes exist in the vector space $H_k(M, \emptyset) = H_k(M)$ and in classical persistent homology are assigned a death parameter of infinity. The object in Fig. 1 illustrates the parameter space Θ and has classes of each type.

Remark 2.5 To preempt any confusion, we note a difference in our nomenclature from some papers, including (Cohen-Steiner et al. 2009b). What we call essential classes above are instead called "extended". We prefer the term "essential" as these classes

do indeed correspond to the essential classes of M. Furthermore it means we can use "extended" to refer to any class in the extended persistence module.

We also partition the elements of the interval decomposition of extended persistent homology into three sets depending on whether they are ordinary, relative or essential. Following (Carlsson et al. 2019) we further split the essential classes into *positive* and *negative* types. For an essential class with birth time (s, Ord) and death time (t, Rel), we say it is positive if $s < t$ and negative if $s > t$.

We can express the extended persistence module as a direct sum of ordinary, relative and essential persistence modules. For an extended persistence module constructed from sublevel and superlevel set filtrations of $f : M \to \mathbb{R}$ denote these submodules by $\text{Ord}_k(M, f)$, $\text{Rel}_k(M, f)$ and $\text{Ess}_k^+(M, f)$ and $\text{Ess}_k^-(M, f)$, which are each persistence modules over $\mathbb{R}$. For $\text{Rel}_k(M, f)$ and $\text{Ess}_k^-(M, f)$ the order of parameters in $\mathbb{R}$ is reversed—that is, the real value associated with the birth time is larger than the real value associated with the death time. Note that in the case of height functions over subsets of $\mathbb{R}^2$ (cf. the example in Fig. 1) Proposition 4.20 implies that $\text{Ess}_0 = \text{Ess}_0^+$ and $\text{Ess}_1 = \text{Ess}_1^-$.

2.2.1 Duality

There is a form of duality between the ordinary persistent homology of $f : M \to \mathbb{R}$ and the relative persistent homology of $(-f) : M \to \mathbb{R}$. This follows from results in De Silva et al. (2011) but that paper uses substantially different notation to us. Furthermore, that paper considers filtrations of simplicial complexes, a context where we cannot naively switch between sublevel and superlevel sets. For these reasons, we rewrite their proposition to suit the requirements of our setting.

Proposition 2.6 (Proposition 2.4 in De Silva et al. (2011)) *Let $\mathbb{M} = \{M_t\}$ be a filtration of simplicial complexes. Let $\text{PH}_k(\mathbb{M})$ be the persistence module of degree-k persistent homology of the filtration $\mathbb{M}$. Let $\text{PH}_k^0(\mathbb{M})$ be the restriction of $\text{PH}_k(\mathbb{M})$ to persistence classes with finite lifetimes. Let $\text{PH}_{k+1}(M_\infty, \mathbb{M})$ be the persistence module of relative homology classes $H_{k+1}(M_\infty, M_t)$ and let $\text{PH}_{k+1}^0(M_\infty, \mathbb{M})$ be the restriction of $\text{PH}_{k+1}(M_\infty, \mathbb{M})$ to persistence classes with finite lifetimes. Then $\text{PH}_k^0(\mathbb{M})$ and $\text{PH}_{k+1}^0(M_\infty, \mathbb{M})$ are isomorphic.*

Corollary 2.7 *Let M be a finite simplicial complex, with vertex set V, and geometric realisation $|M|$. Let $f : |M| \to \mathbb{R}$ be a continuous map such that on each cell f is the linear interpolation of the values on its vertices. We have a bijection ρ between the interval modules in the interval decomposition of $\text{Ord}_k(M, f)$ to that of $\text{Rel}_{k+1}(M, (-f))$ with*

$$\rho(\mathcal{I}_{[(b,\text{ord}),(d,\text{ord}))}) = \mathcal{I}_{[(-b,\text{rel}),(-d,\text{rel}))}$$

Proof The $\text{PH}_{k+1}^0(M_\infty, M_t)$ of De Silva et al. (2011) is the relative homology of M with respect to the (increasing t) sequence $M_t = f^{-1}(-\infty, t]$. But $f^{-1}(-\infty, t] = (-f)^{-1}[-t, \infty)$, so the sequence M_t of sublevel sets of f is identical to a sequence of

superlevel sets, M^s, of $(-f)$, with $s = -t$. Note that when the filtration is expressed as superlevel sets of $(-f)$, the parameter s is a decreasing one, as used in the relative part of an extended persistence module.

From Proposition 2.6, we have a bijection between the intervals in the interval decompositions with $\mathcal{I}_{[b,d)} \subset \mathrm{PH}_k^0(\mathbb{M})$ matched to $\mathcal{I}_{[b,d)} \subset \mathrm{PH}_{k+1}^0(M, \mathbb{M})$. Composing this with the reparameterisation to superlevel set notation we have $\mathcal{I}_{[b,d)} \subset \mathrm{PH}_{k+1}^0(M, \mathbb{M})$ rewritten as $\mathcal{I}_{[(-b,\mathrm{rel}),(-d,\mathrm{rel}))} \subset \mathrm{Rel}(M, (-f))$. □

We note that this duality result is quite different from the duality theorem of Cohen-Steiner et al. (2009b), which is proved in the case that M is a triangulated d-manifold. That paper goes on to also establish a symmetry theorem for extended persistence for functions over manifolds without boundary, which we discuss in our notation and context below.

2.2.2 Symmetry

In the case that M is a manifold we find that the information content in extended persistence modules is greatly reduced by the isomorphisms established in the following result.

Proposition 2.8 (Symmetry theorem of Cohen-Steiner et al. (2009b)) *Let M be a triangulated d-manifold and $f : M \to \mathbb{R}$ be a piecewise-linear function interpolating the values on the vertices of M. There are bijections, $\psi_\bullet$, between submodules of extended persistence for f and $(-f)$ as follows:*

$$\psi_O : \mathrm{Ord}_k(M, f) \to \mathrm{Ord}_{d-k-1}(M, -f) \quad \mathcal{I}_{[(b,\mathrm{ord}),(d,\mathrm{ord}))} \mapsto \mathcal{I}_{[(-d,\mathrm{ord}),(-b,\mathrm{ord}))}$$
$$\psi_E : \mathrm{Ess}_k(M, f) \to \mathrm{Ess}_{d-k}(M, -f) \quad\;\; \mathcal{I}_{[(b,\mathrm{ord}),(d,\mathrm{rel}))} \mapsto \mathcal{I}_{[(-d,\mathrm{ord}),(-b,\mathrm{rel}))}$$
$$\psi_R : \mathrm{Rel}_k(M, f) \to \mathrm{Rel}_{d-k+1}(M, -f) \quad \mathcal{I}_{[(b,\mathrm{rel}),(d,\mathrm{rel}))} \mapsto \mathcal{I}_{[(-d,\mathrm{rel}),(-b,\mathrm{rel}))}$$

Remark 2.9 We note that Cohen-Steiner et al. (2009b) has a typographical error in the dimensions for the relative homology classes, that was corrected in Cohen-Steiner et al. (2009a).

Proof As in Cohen-Steiner et al. (2009b), first use Lefschetz duality $H_k(X, \partial X) \leftrightarrow H_{d-k}(X, \emptyset)$ with $X = M_t$ and the excision theorem to see that

$$H_k(M, M^t) = H_k(M_t, \partial M_t) = H_{d-k}(M_t, \emptyset).$$

Combined with the inclusion-induced maps on homology, this gives a bijection between the finite intervals of ordinary and relative homology in complementary dimensions: $\mathrm{Rel}_k(M, f) \leftrightarrow \mathrm{Ord}_{d-k}(M, f)$, with $\mathcal{I}_{[(b,\mathrm{rel}),(d,\mathrm{rel}))} \mapsto \mathcal{I}_{[(d,\mathrm{ord}),(b,\mathrm{ord}))}$. The same relationship holds for the essential homology classes: $\mathrm{Ess}_k(M, f) \leftrightarrow \mathrm{Ess}_{d-k}(M, f)$, with $\mathcal{I}_{[(b,\mathrm{ord}),(d,\mathrm{rel}))} \mapsto \mathcal{I}_{[(d,\mathrm{ord}),(b,\mathrm{rel}))}$. Note these bijections are those established by the duality theorem of Cohen-Steiner et al. (2009b). Combined with the duality result 2.7 above, we now see that

$$\mathrm{Rel}_k(M, f) \leftrightarrow \mathrm{Ord}_{d-k}(M, f) \leftrightarrow \mathrm{Rel}_{d-k+1}(M, -f)$$

$$\mathrm{Ess}_k(M, f) \leftrightarrow \mathrm{Ess}_{d-k}(M, f) \leftrightarrow \mathrm{Ess}_{d-k}(M, -f)$$
$$\mathrm{Ord}_k(M, f) \leftrightarrow \mathrm{Rel}_{d-k}(M, f) \leftrightarrow \mathrm{Ord}_{d-k-1}(M, -f)$$

Composing the two bijections establishes the maps $\psi_\bullet$ in each case. $\qquad\square$

Remark 2.10 Our application to binary images has data M that are manifolds with boundary, so the duality and symmetry theorems of Cohen-Steiner et al. (2009b) do not apply directly. We use the duality result of De Silva et al. (2011) (expressed as Corollary 2.7) to reduce the number of directions required when computing the extended persistent homology transform, since it gives a bijection between the intervals for height filtrations in opposite directions. Since the boundary ∂M of a manifold with boundary $(M, \partial M)$ is a manifold we use the above symmetry result to characterise the essential classes of a height filtration in $\mathrm{Ess}_0(\partial M, f)$ and $\mathrm{Ess}_{n-1}(\partial M, f)$ in Proposition 4.20.

3 Wasserstein distance between extended persistence modules

3.1 Wasserstein distances between persistence modules

There are many possible metrics between persistence modules, and various representations of them. In this paper we restrict our attention to Wasserstein distances. Wasserstein distances between persistence modules are usually defined in terms of the points in their corresponding persistence diagrams. However, given our desire to study extended persistence, we rephrase the definitions here in terms of persistence modules over a totally ordered set. Wasserstein distances are a form of optimal transport metric. A transportation plan between two persistence modules matches subsets of intervals from each, with the remaining unmatched intervals paired with ephemeral intervals. Since every persistence module considered in this paper is isomorphic to a direct sum of interval modules it is sufficient to define our transportation plans between persistence modules written in this form.

Definition 3.1 Let Θ be a totally ordered set and $\mathcal{P} = \bigoplus_{I_i \in S^{\mathcal{P}}} \mathcal{I}_{I_i}$ and $\mathcal{Q} = \bigoplus_{I_j \in S^{\mathcal{Q}}} \mathcal{I}_{I_j}$ persistence modules over Θ. A *transportation plan* between $\mathcal{P}$ and $\mathcal{Q}$ is a triple $T = (\hat{S}^{\mathcal{P}}, \hat{S}^{\mathcal{Q}}, \rho)$ where $\hat{S}^{\mathcal{P}} \subset S^{\mathcal{P}}$, $\hat{S}^{\mathcal{Q}} \subset S^{\mathcal{Q}}$ and $\rho : \hat{S}^{\mathcal{P}} \to \hat{S}^{\mathcal{Q}}$ is a bijection. We call the intervals in $\hat{S}^{\mathcal{P}}$ and in $\hat{S}^{\mathcal{Q}}$ *matched* intervals in T, and we call the intervals in $S^{\mathcal{P}} \backslash \hat{S}^{\mathcal{P}}$ and in $S^{\mathcal{Q}} \backslash \hat{S}^{\mathcal{Q}}$ *unmatched* intervals in T.

Each transportation plan has an associated cost, constructed analogously to an L^p function metric. This in turn depends on the metric used to measure distance between points in Θ, which we define below.

Definition 3.2 We call $(\Theta, \leq, \mathrm{dist})$ a *totally ordered metric space* if $(\Theta, \leq)$ is a totally ordered set, and dist is an extended metric over Θ such that $\mathrm{dist}(\beta, \gamma) \leq \mathrm{dist}(\alpha, \gamma)$ and $\mathrm{dist}(\alpha, \beta) \leq \mathrm{dist}(\alpha, \gamma)$ whenever $\alpha \leq \beta \leq \gamma$.

From the metric on Θ we obtain a p-distance between intervals over Θ, analogous to the l^p distance between points in $\mathbb{R}^2$. Given two intervals $\mathcal{I}$ and $\mathcal{I}'$, the p-distance

(for $p \in [1, \infty)$) is defined as

$$\text{dist}_p(\mathcal{I}, \mathcal{I}') = \big(\text{dist}(\mathfrak{b}(\mathcal{I}), \mathfrak{b}(\mathcal{I}'))^p + \text{dist}(\mathfrak{d}(\mathcal{I}), \mathfrak{d}(\mathcal{I}'))^p\big)^{1/p}.$$

The bottleneck, or ∞-distance, between intervals is

$$\text{dist}_\infty(\mathcal{I}, \mathcal{I}') = \max\{\text{dist}(\mathfrak{b}(\mathcal{I}), \mathfrak{b}(\mathcal{I}')), \text{dist}(\mathfrak{d}(\mathcal{I}), \mathfrak{d}(\mathcal{I}'))\}.$$

Note that for general interval modules this is actually a pseudo-distance as it cannot distinguish between intervals with open or closed endpoints. However, if the persistence modules are constructed from filtrations involving closed sublevel and superlevel sets then the intervals are always half-open, including the birth parameter and not including the death parameter. When restricted to such half-open interval modules the above definition of dist$_p$ will satisfy the identity of indiscernibles, making it an actual distance. Throughout this paper we will work exclusively with persistence modules that have these half-open intervals.

The final ingredient we need before defining the transportation plans and their costs is the notion of an "empty interval". For persistence diagrams these are points on the diagonal, corresponding to intervals of zero length in the usual setting of persistence modules over $\mathbb{R}$. In the general definition of Wasserstein distance we are allowed to fix any subset of interval modules to perform this role. We call this set the *ephemeral intervals* denoted Eph. This name is inspired by the definition of an ephemeral persistence module as one with distance zero to the trivial persistence module (see Chazal et al. 2014).

We now define the cost of a transportation plan using p-distances between intervals which are matched, and to each unmatched interval we assign a cost which is the distance to its closest ephemeral interval.

Definition 3.3 Let Θ be a totally ordered set; $\mathcal{P} = \bigoplus_{a \in S^\mathcal{P}} \mathcal{I}_a$ and $\mathcal{Q} = \bigoplus_{b \in S^\mathcal{Q}} \mathcal{I}_b$ be persistence modules over the ordered metric space $(\Theta, \leq, \text{dist})$. Let Eph denote the set ephemeral intervals over Θ. Let $T = (\hat{S}^\mathcal{P}, \hat{S}^\mathcal{Q}, \rho)$ be a transportation plan between $\mathcal{P}$ and $\mathcal{Q}$. For $p \in [1, \infty)$ we define the p-cost of T by

$$c_p(T)^p = \sum_{a \in \hat{S}^\mathcal{P}} \text{dist}_p(\mathcal{I}_a, \mathcal{I}_{\rho(a)})^p + \sum_{a \in S^\mathcal{P} \setminus \hat{S}^\mathcal{P}} \inf_{\mathcal{I} \in \text{Eph}} \{\text{dist}_p(\mathcal{I}_a, \mathcal{I})^p\}$$

$$+ \sum_{b \in S^\mathcal{Q} \setminus \hat{S}^\mathcal{Q}} \inf_{\mathcal{I} \in \text{Eph}} \{\text{dist}_p(\mathcal{I}_b, \mathcal{I})^p\}$$

and

$$c_\infty(T) = \max\Big\{ \sup_{a \in \hat{S}^\mathcal{P}} \{\text{dist}_\infty(\mathcal{I}_a, \mathcal{I}_{\rho(a)})\}, \sup_{a \in S^\mathcal{P} \setminus \hat{S}^\mathcal{P}} \{\inf_{\mathcal{I} \in \text{Eph}} \{\text{dist}_\infty(\mathcal{I}_a, \mathcal{I})\}\},$$

$$\sup_{b \in S^\mathcal{Q} \setminus \hat{S}^\mathcal{Q}} \{\inf_{\mathcal{I} \in \text{Eph}} \{\text{dist}_\infty(\mathcal{I}_b, \mathcal{I})\}\}\Big\}$$

Observe that $c_\infty(T)$ is the limit of $c_p(T)$ as p goes to infinity. The Wasserstein distance is defined as the infimum of the costs of all transportation plans. Note that there is always at least one possible transportation plan as we can choose $\hat{S}^\mathcal{P}$ and $\hat{S}^\mathcal{Q}$ to be empty.

Definition 3.4 Fix $p \in [1, \infty)$. Let Θ be a totally ordered set and $\mathcal{P} = \bigoplus_{a \in S^\mathcal{P}} \mathcal{I}_a$ and $\mathcal{Q} = \bigoplus_{b \in S^\mathcal{Q}} \mathcal{I}_b$ be persistence modules over the ordered metric space $(\Theta, \leq, \text{dist})$. The *p-Wasserstein* distance between $\mathcal{P}$ and $\mathcal{Q}$ is

$$W_p(\mathcal{P}, \mathcal{Q}) = \inf\{c_p(T) \mid T \text{ a transportation plan between } \mathcal{P} \text{ and } \mathcal{Q}\}.$$

The *bottleneck* distance between $\mathcal{P}$ and $\mathcal{Q}$ is

$$W_\infty(\mathcal{P}, \mathcal{Q}) = \inf\{c_\infty(T) \mid T \text{ a transportation plan between } \mathcal{P} \text{ and } \mathcal{Q}\}.$$

This definition agrees with the standard definitions of Wasserstein and bottleneck distances between persistence diagrams when Θ is the real line with its standard order, $\text{dist}(s, t) = |s - t|$, and $\text{Eph} = \{[t, t] : t \in \mathbb{R}\}$. More generally, for any totally ordered metric space and any choice for the set of ephemeral intervals, the Wasserstein distance defined above will determine an extended metric. Again, for general persistence modules this will be, strictly speaking, a pseudo-distance. But, as discussed earlier, in this paper the persistence modules will only contain appropriate half-open intervals and $W_p(\mathcal{P}, \mathcal{Q})$ satisfies the identity of indiscernibles.

3.2 Wasserstein distance for extended persistence

The Wasserstein distance between persistence modules is specified by the ordered metric space and set of ephemeral interval modules. Recall from Sect. 2.2 that extended persistent modules have parameter set $\Theta = O \cup R$, with $O = \{(t, \text{Ord}) : t \in \mathbb{R}\}$ and $R = \{(t, \text{Rel}) : t \in \mathbb{R}\}$, and the total order over P is

$$(s, \text{Ord}) < (t, \text{Ord}) \text{ when } s < t$$
$$(s, \text{Rel}) < (t, \text{Rel}) \text{ when } s > t$$
$$(s, \text{Ord}) < (t, \text{Rel}) \text{ for all } s, t.$$

We make Θ an ordered metric space by constructing an appropriate extended metric over Θ. A natural choice is

$$\text{dist}((s, \text{Ord}), (t, \text{Ord})) = |s - t| \text{ for all } s, t.$$
$$\text{dist}((s, \text{Rel}), (t, \text{Rel})) = |s - t| \text{ for all } s, t.$$
$$\text{dist}((s, \text{Ord}), (t, \text{Rel})) = \infty \text{ for all } s, t.$$

We also need to define the set of ephemeral interval modules; there are three different types: ordinary, relative and essential. We set

$$\text{Eph} = \{\mathcal{I}_{[(t,\text{Ord}),(t,\text{Ord}))} : t \in \mathbb{R}\} \cup \{\mathcal{I}_{[(t,\text{Rel}),(t,\text{Rel}))} : t \in \mathbb{R}\}$$
$$\cup \{\mathcal{I} : \mathfrak{b}(\mathcal{I}) = (t, \text{Ord}) \text{ and } \mathfrak{d}(\mathcal{I}) = (t, \text{Rel}) \text{ for some } t \in \mathbb{R}\}.$$

Fig. 2 Optimal transport between the degree-1 essential classes in the pretzel (three non-trivial degree-1 essential classes) and the donut (one non-trivial degree-1 essential class). The height function is the horizontal coordinate. The optimal matching pairs the green essential classes to each other. The red and blue essential classes of the pretzel are matched to ephemeral classes for the donut—these are zero length with height parameter located at the mid-point of birth and death parameters for the corresponding essential class of the pretzel. That is, we match $[(b_i, \mathrm{Ord}), (d_i, \mathrm{Rel}))$ to $[(\frac{b_i+d_i}{2}, \mathrm{Ord}), (\frac{b_i+d_i}{2}, \mathrm{Rel}))$ (color figure online)
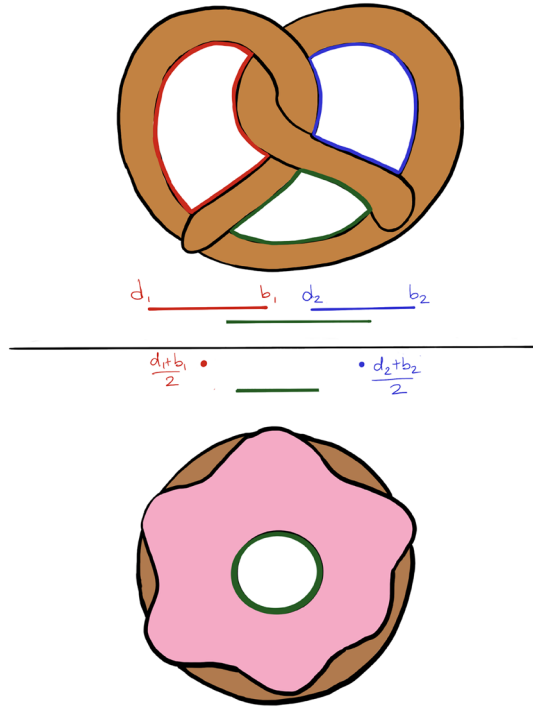
An example that illustrates how essential classes are paired with ephemeral classes is shown in Fig. 9.

For computational purposes it is much easier to split the calculation of distances between extended persistence modules into separate calculations for the submodules of the types Ord, Rel, Ess$^+$ and Ess$^-$. This is justified by the following proposition.

Proposition 3.5 *Let $\mathcal{P}$ and $\mathcal{Q}$ be extended persistence modules in a single homology degree and let $\mathcal{P} = \mathrm{Ord}(\mathcal{P}) \oplus \mathrm{Rel}(\mathcal{P}) \oplus \mathrm{Ess}^+(\mathcal{P}) \oplus \mathrm{Ess}^-(\mathcal{P})$ and $\mathcal{Q} = \mathrm{Ord}(\mathcal{Q}) \oplus \mathrm{Rel}(\mathcal{Q}) \oplus \mathrm{Ess}^+(\mathcal{Q}) \oplus \mathrm{Ess}^-(\mathcal{Q})$ be their decomposition into the four types of classes. Then*

$$
\begin{aligned}
W_p(\mathcal{P}, \mathcal{Q})^p =& W_p(\mathrm{Ord}(\mathcal{P}), \mathrm{Ord}(\mathcal{Q}))^p + W_p(\mathrm{Rel}(\mathcal{P}), \mathrm{Rel}(\mathcal{Q}))^p \\
& + W_p(\mathrm{Ess}^-(\mathcal{P}), \mathrm{Ess}^-(\mathcal{Q}))^p + W_p(\mathrm{Ess}^+(\mathcal{P}), \mathrm{Ess}^+(\mathcal{Q}))^p
\end{aligned}
$$

for $p \in [1, \infty)$ and

$$
\begin{aligned}
W_\infty(\mathcal{P}, \mathcal{Q}) = \max \Big\{ & W_\infty(\mathrm{Ord}(\mathcal{P}), \mathrm{Ord}(\mathcal{Q})), W_\infty(\mathrm{Rel}(\mathcal{P}), \mathrm{Rel}(\mathcal{Q})), \\
& W_\infty(\mathrm{Ess}^-(\mathcal{P}), \mathrm{Ess}^-(\mathcal{Q})), W_\infty(\mathrm{Ess}^+(\mathcal{P}), \mathrm{Ess}^+(\mathcal{Q})) \Big\}.
\end{aligned}
$$

Proof The right hand side of the both equations is the infimum of transportation costs over the set of transportation plans which never match any intervals of different types.

It is thus sufficient to show that for any transportation plan between $\mathcal{P}$ and $\mathcal{Q}$ there is another transportation T with the same or lesser cost such that any matched pair within T keeps to the same type. Any two intervals of different types of Rel, Ord or Ess are an infinite distance apart. Since every interval module has finite distance to some ephemeral interval it will always be more efficient to change any interval that is matched to a different type to instead be unmatched. Similarly there is a higher cost to match positive with negative essential classes than to leave both unmatched. □

It is worth observing that in previous work, such as Carlsson et al. (2019), Bauer et al. (2020), the extended persistent homology modules are represented by multiple persistence diagrams, separating the different types into their own persistence diagrams. The ordinary persistence diagram has points above the diagonal, the relative persistence diagram has points only below the diagonal, and the essential persistence diagram has points on both sides—positive above and negative below. The bottleneck distance in Bauer et al. (2020) is then defined as the formula within Proposition 3.5.

Remark 3.6 We believe that the Wasserstein distance could also be defined analogous to the algebraic Wasserstein distance in Skraba and Turner (2020) but adapted to extended persistence, and that these two versions of Wasserstein distances would be equivalent. Given the enormous homological algebra set up required to prove such a result it is beyond the scope of this paper and left as a future direction of research.

4 Morse theory for manifolds with boundary and extended persistence

This section contains the main theoretical results relating extended persistence of a height function over a manifold with boundary to that of the same function restricted to the boundary. We establish these results using Morse theory, a standard technique when working with persistence modules built from sublevel set filtrations. Previous results, however, apply only to functions on manifolds, and not to functions on manifolds with boundary. The presence of a boundary requires extra analysis to characterise critical points located on this boundary. We start by summarising the necessary definitions and results from Morse theory covering both the smooth and piecewise-linear settings.

4.1 Background: smooth and PL Morse theory

We need our results about extended persistent homology to hold for both the smooth (theoretical) case, and the piecewise-linear setting relevant to numerical computations. Most of the theorems and their proofs are effectively the same but we must first set up the definitions and relevant lemmas about critical points. The background theory is covered for the smooth case in Braess (1974), Jankowski and Rubinsztejn (1972), and the piecewise linear case in Grunert et al. (2019). We direct readers interested in more details to these references.

Although regular and critical points and their indices in Morse theory are more commonly defined in terms of the derivatives and Hessian of a function, this approach

does not translate well to the PL setting. There is, however, an equivalent approach to defining critical points and indices that uses polynomial functions over charts, and this can be easily adapted to the PL setting. To make this paper self-contained we start by recalling the definitions of smooth and PL manifold (with or without boundary) in terms of charts.

Definition 4.1 For a topological space, M, and an open subset $U \subset M$, a *chart* is a homeomorphism $\phi : U \to \phi(U)$ where $\phi(U)$ is a subset of Euclidean space. An *atlas* for M is an indexed family of charts $\{(U_\alpha, \phi_\alpha)\}$ that cover M, i.e., $\cup U_\alpha = M$. A *topological n-manifold* is a second countable, Hausdorff space equipped with an atlas where the codomain of each ϕ_α is an open subset of $\mathbb{R}^n$. A *topological n-manifold with boundary* is a second countable, Hausdorff space equipped with an atlas where the codomain of each ϕ_α is an open subset of $[0, \infty) \times \mathbb{R}^{n-1}$.

To introduce the adjectives smooth and piecewise linear (PL) we need to discuss the compatibility of ϕ_α and ϕ_β on the intersections of their domains. Given two charts (U_α, ϕ_α) and (U_β, ϕ_β) where $U_\alpha \cap U_\beta$ has non-empty intersection we can define two different maps by restricting the domains of ϕ_α and ϕ_β to $U_\alpha \cap U_\beta$. The new homeomorphisms are $\phi_\alpha \circ (\phi_\beta)^{-1} : \phi_\beta(U_\alpha \cap U_\beta) \to \phi_\alpha(U_\alpha \cap U_\beta)$ and $\phi_\beta \circ (\phi_\alpha)^{-1} : \phi_\alpha(U_\alpha \cap U_\beta) \to \phi_\beta(U_\alpha \cap U_\beta)$. These are called the *transition maps* between charts.

Definition 4.2 A topological n-manifold, with or without boundary, is called *smooth* if its transition maps are smooth. It is called *piecewise-linear* (PL for short) if its transition maps are piecewise-linear.

We say that $\{(U_\alpha, \phi_\alpha)\}$ is *maximal* if there does not exist another atlas containing it with more charts. A maximal atlas is often referred to as the smooth structure, or respectively, the PL structure of a manifold. Once we have a smooth (or PL) structure we can define what it means for a function $f : M \to \mathbb{R}$ to be smooth or piecewise linear.

Definition 4.3 Let M be a smooth n-manifold, with or without boundary, with smooth (respectively PL) structure $\{(U_\alpha, \phi_\alpha)\}$. A function $f : M \to \mathbb{R}$ is *smooth* (respectively *PL*) if $\phi_\alpha^{-1} \circ f : \phi_\alpha(U_\alpha) \to \mathbb{R}$ is smooth (respectively *PL*) for all charts (U_α, ϕ_α).

An example to keep in mind is M being a smooth or piecewise linear n-dimensional subset of $\mathbb{R}^d$ with its structure inherited from the embedding. A simple function on such a manifold is the height function with respect to some unit vector $v \in S^{d-1}$, i.e., $f(x) = v \cdot x$.

The classical approach to defining critical points in Morse theory is as follows. For a manifold M without boundary and a smooth function $f : M \to \mathbb{R}$. Let $p \in M$ and choose a chart (U, ϕ) with $p \in U$. We say that $p \in M$ is a *critical point* of f if $d(f \circ \phi^{-1})(\phi(p)) = 0$. A critical point is *non-degenerate* if the Hessian of $f \circ \phi$ at p is non-singular. We then say the *Morse index of f at p* is the number of negative eigenvalues of the Hessian, counting multiplicity. A point is *regular* if it is not critical. These definitions are well defined as they do not depend on the choice of chart (see Milnor 1963).

Instead of using definitions for critical and regular points in terms of the derivative, we need an alternative that will be more adaptable to the PL setting. By using the implicit function theorem we can redefine regular points by the existence of a linear function over some chart. We can also remove the need to reference the Hessian for defining the index of a critical point by using the Morse Lemma.

Lemma 4.4 (Morse Lemma) *Let M be a smooth n-manifold without boundary and $f : M \to \mathbb{R}$ a smooth function. The point $p \in M$ is a regular point of f if and only if there is a chart (U, ϕ) where $\phi(p) = 0$ and*

$$f \circ \phi^{-1}(x_1, x_2, \ldots x_n) = f(p) + x_n$$

in some neighbourhood of 0.

The point $p \in M$ is a non-degenerate critical point of f with Morse index k if and only if there is a chart (U, ϕ) where $\phi(p) = 0$ and

$$f \circ \phi^{-1}(x_1, x_2, \ldots x_n) = f(p) - (x_1)^2 - (x_2)^2 - \ldots - (x_k)^2 + (x_{k+1})^2 + \ldots + (x_n)^2$$

in a neighbourhood of 0.

The proof of this lemma is covered in Milnor (1963). We use it as an equivalent definition of a regular point and a non-degenerate critical point of Morse index k. In the piecewise linear setting the only modification is to replace squares with absolute values.

Definition 4.5 Let M be an n-dimensional PL manifold without boundary and $f : M \to \mathbb{R}$ a PL function. The point $p \in M$ is a *regular* point of f if and only if there is a chart (U, ϕ) containing p of the form

$$f \circ \phi^{-1}(x_1, x_2, \ldots x_n) = f(p) + x_n.$$

The point $p \in M$ is a non-degenerate critical point of f with Morse index k if and only if there is a chart (U, ϕ) with $\phi(p)$ which is of the form

$$f \circ \phi^{-1}(x_1, x_2, \ldots x_n) = f(p) - |x_1| - |x_2| - \ldots - |x_k| + |x_{k+1}| + \ldots + |x_n|$$

in a neighbourhood of 0.

We now need to generalise the definitions of regular and critical points to the case of a function over a manifold with boundary $(M, \partial M)$. Points in the interior of M are treated exactly as above, so we need only discuss the case for points on the boundary. We again phrase the definitions using charts to make it easy to move between smooth and PL settings, following the terminology and notation in Grunert et al. (2019). Recall that a chart containing a point, $p \in \partial M$ is homeomorphic to a subset of $\{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \mid x_1 \geq 0\}$, with $\phi(p) = (0, x_2, \ldots, x_n)$.

Definition 4.6 Let $(M, \partial M)$ be a smooth (respectively PL) n-manifold with boundary and $f : M \rightarrow \mathbb{R}$ a smooth (respectively PL) function. The point $p \in \partial M$ is a *regular* point of f if and only if there is a chart (U, ϕ) with $\phi(p) = 0$ of the form $f \circ \phi^{-1}(x_1, x_2, \ldots x_n) = f(p) + x_n$.

A point on the boundary is critical if it is critical for f restricted to ∂M, but the definition of its index requires additional information about whether the function increases or decreases as we move into the manifold.

Definition 4.7 Let $(M, \partial M)$ be a smooth n-manifold with boundary and $f : M \rightarrow \mathbb{R}$ a smooth function. The point $p \in \partial M$ is a *non-degenerate critical* point of f with index (k, η) if only if there is a chart (U, ϕ) with $\phi(p) = 0$ such that

$$
\begin{aligned}
f \circ \phi^{-1}(x_1, x_2, \ldots x_n) &= f(p) + \eta x_1 - (x_2)^2 \\
&- \ldots - (x_{k+1})^2 + (x_{k+2})^2 + \ldots + (x_n)^2.
\end{aligned}
$$

The second term of the index, $\eta \in \{-1, 1\}$, defines the sign of the critical point: if $\eta = 1$ we say that p is $(+)$-critical, and if $\eta = -1$, then p is $(-)$-critical.

The analogous definition for the piecewise linear case is:

Definition 4.8 Let $(M, \partial M)$ be a PL n-manifold with boundary and $f : M \rightarrow \mathbb{R}$ a PL function. The point $p \in \partial M$ is a *non-degenerate critical* point of f of index (k, η) if there is a chart (U, ϕ) with $\phi(p) = 0$ of the form

$$
f \circ \phi^{-1}(x_1, x_2, \ldots x_n) = f(p) + \eta x_1 - |x_2| - \ldots - |x_{k+1}| + |x_{k+2}| + \ldots + |x_n|.
$$

Again, p is $(+)$-critical when $\eta = +1$ and $(-)$-critical when $\eta = -1$.

Please note that there is inconsistency within the literature in terms of sign conventions for critical points on the boundary and our choice may differ from sources the reader is familiar with.

Now we have the definitions for all the different types of critical point, we can define what a Morse function is for both the smooth and PL settings.

Definition 4.9 Given a smooth (respectively PL) manifold with boundary $(M, \partial M)$, we say that $f : M \rightarrow \mathbb{R}$ is a *Morse function* if

- f is smooth (respectively PL)
- None of the critical points of $f|_{\mathrm{int}(M)}$ and $f|_{\partial M}$ are degenerate.
- All the critical values for $f|_{\mathrm{int}(M)}$ and $f|_{\partial M}$ combined are distinct and finite in number.

In the following we describe the (persistent) homology in terms of the signs of critical points so it is useful to have notation for this.

Definition 4.10 Suppose $f : (M, \partial M) \rightarrow \mathbb{R}$ is a Morse function. Let $\mathrm{Crit}(f, k)$ denote the set of index-k critical points of f; these points must lie in the interior of M. Let $\mathrm{Crit}(f, (k, \eta))$ denote the set of critical points of $f|_{\partial M}$ with index (k, η). If $p \in \partial M$ is a critical point of $f|_{\partial M}$, with index (k, η) denote the sign of p by $\mathrm{sgn}(f, p) = \eta$.

Highly analogous to the well-known theory of Morse functions on manifolds, we can use the index of critical points to compute the relative homology of nearby sublevel sets of f.

Proposition 4.11 *Let $(M, \partial M)$ be a smooth (respectively PL) manifold with boundary and $f : M \to \mathbb{R}$ a smooth (respectively PL) Morse function. We consider homology with coefficients in a field $\mathbb{F}$, and use Kronecker delta notation δ_i^k below.*

- *If t is not a critical value of either f or $f|_{\partial M}$ then $H_i(M_{t+\epsilon}, M_{t-\epsilon}) = 0$ for all i and all $\epsilon > 0$ sufficiently small.*
- *If $p \in \mathrm{Crit}(f, k)$ then $H_i(M_{f(p)+\epsilon}, M_{f(p)-\epsilon}) = \delta_i^k \mathbb{F}$ for all i and for all $\epsilon > 0$ sufficiently small.*
- *If $p \in \mathrm{Crit}(f, (k, -1))$ then $H_i(M_{f(p)+\epsilon}, M_{f(p)-\epsilon}) = 0$ for $\epsilon > 0$ sufficiently small.*
- *If $p \in \mathrm{Crit}(f, (k, +1))$ then $H_i(M_{f(p)+\epsilon}, M_{f(p)-\epsilon}) = \delta_i^k \mathbb{F}$ for all i, for $\epsilon > 0$ sufficiently small.*

For the smooth case, this proposition is proved in Braess (1974) and in Jankowski and Rubinsztejn (1972). Please note that in Jankowski and Rubinsztejn (1972) they use the term "m-function" for Morse function. Some minor massaging is needed to convert their results to the homology statements above as they describe the changes in terms of glueing cells. The PL version of this proposition is proved in Grunert et al. (2019).

We can determine critical points and indices of $(-f)$ from those of f using charts, as summarised in the following lemma which holds for both the smooth and PL settings.

Lemma 4.12 *Let $(M, \partial M)$ be an n-manifold with boundary and $f : M \to \mathbb{R}$ a Morse function. Then $(-f) : M \to \mathbb{R}$ is also a Morse function with $\mathrm{Crit}((-f), k) = \mathrm{Crit}(f, n - k)$ and $\mathrm{Crit}((-f), (k, \eta)) = \mathrm{Crit}(f, (n - k - 1, -\eta))$ for $\eta = \pm 1$.*

This facilitates analogous homology results as in Proposition 4.11 but for relative homology of superlevel sets.

Corollary 4.13 *Let $(M, \partial M)$ be a smooth (respectively PL) n-manifold with boundary and $f : M \to \mathbb{R}$ a smooth (respectively PL) Morse function.*

- *If t is not a critical value of either f or $f|_{\partial M}$ then $H_i(M^{t-\epsilon}, M^{t+\epsilon}) = 0$ for all i and all $\epsilon > 0$ sufficiently small.*
- *If $p \in \mathrm{Crit}(f, n - k)$ then $H_i(M^{f(p)-\epsilon}, M^{f(p)+\epsilon}) = \delta_i^k \mathbb{F}$ for all i and for all $\epsilon > 0$ sufficiently small.*
- *If $p \in \mathrm{Crit}(f, (n - k - 1, +1))$ then $H_i(M^{f(p)-\epsilon}, M^{f(p)+\epsilon}) = 0$ for all i and for $\epsilon > 0$ sufficiently small.*
- *If $p \in \mathrm{Crit}(f, (n - k - 1, -1))$ then $H_i(M^{f(p)-\epsilon}, M^{f(p)+\epsilon}) = \delta_i^k \mathbb{F}$ for all i and for $\epsilon > 0$ sufficiently small.*

Proof We first want to write the superlevel sets of f in terms of sublevel sets of $(-f)$. We have $M^s = (-f)^{-1}(-\infty, -s]$, and thus

$$H_i(M^{f(p)-\epsilon}, M^{f(p)+\epsilon})$$
$$= H_i\big((-f)^{-1}(-\infty, (-f)(p) + \epsilon], (-f)^{-1}(-\infty, (-f)(p) - \epsilon]\big).$$

If t is not a critical value for f nor $f|_{\partial M}$ then by Lemma 4.12 $(-t)$ is not a critical value of $(-f)$ nor $(-f)|_{\partial M}$. By Proposition 4.11 we know

$$H_i((-f)^{-1}(-\infty, -t + \epsilon], (-f)^{-1}(-\infty, -t - \epsilon]) = 0$$

for all i and all $\epsilon > 0$ sufficiently small.

If $p \in \mathrm{Crit}(f, n - k)$, then by Lemma 4.12 we have $p \in \mathrm{Crit}((-f), k)$. If $p \in \mathrm{Crit}(f, (n - k - 1, -1))$ then by Lemma 4.12 $p \in \mathrm{Crit}((-f), (k, +1))$. In both cases we can apply Proposition 4.11, with $(-f)$ at p, which implies that

$$H_i\big((-f)^{-1}(-\infty, (-f)(p) + \epsilon], (-f)^{-1}(-\infty, (-f)(p) - \epsilon]\big) = \delta_i^k \, \mathbb{F}.$$

If $p \in \mathrm{Crit}(f, (n - k - 1, +1))$, then by Lemma 4.12 we have $p \in \mathrm{Crit}((-f), (k, -1))$. By Proposition 4.11, with $(-f)$ at p, we know

$$H_i\big((-f)^{-1}(-\infty, (-f)(p) + \epsilon], (-f)^{-1}(-\infty, (-f)(p) - \epsilon]\big) = 0.$$

for $\epsilon > 0$ sufficiently small. $\qquad\square$

As might be expected, there is a direct relationship between the critical values of Morse functions and the endpoints of intervals in the barcode decomposition of extended persistent homology. We will need to distinguish between endpoints lying in the ordinary and relative parameter spaces as they behave differently.

Let $\mathrm{XPH}(M, f)$ be the extended persistence module constructed from $f : M \to \mathbb{R}$. To ease notation let

$$\mathfrak{b}_k^{\mathrm{ord}}(M, f) := \mathfrak{b}(\mathrm{Ord}_k(M, f) \oplus \mathrm{Ess}_k(M, f))$$

and

$$\mathfrak{b}_k^{\mathrm{rel}}(M, f) := \mathfrak{b}(\mathrm{Rel}_k(M, f))$$

These are the sets of parameters $\{(t, \mathrm{ord})\}$ and $\{(t, \mathrm{rel})\}$ respectively where a new interval begins in the interval decomposition of $\mathrm{XPH}(M, f)$. Similarly let

$$\mathfrak{d}_k^{\mathrm{ord}}(M, f) := \mathfrak{d}(\mathrm{Ord}_k(M, f))$$

and

$$\mathfrak{d}_k^{\mathrm{rel}}(M, f) := \mathfrak{d}(\mathrm{Rel}_k(M, f) \oplus \mathrm{Ess}_k(M, f)).$$

These are the sets of parameters $\{(t, \mathrm{ord})\}$ and $\{(t, \mathrm{rel})\}$ respectively where an interval finishes in the interval decomposition of $\mathrm{XPH}(M, f)$. Furthermore let $\mathfrak{b}_k(M, f) =$

$\mathfrak{b}_k^{\mathrm{ord}}(M, f) \cup \mathfrak{b}_k^{\mathrm{rel}}(M, f)$ and $\mathfrak{d}_k(M, f) = \mathfrak{d}_k^{\mathrm{ord}}(M, f) \cup \mathfrak{d}_k^{\mathrm{rel}}(M, f)$ denote the sets of birth and death parameters respectively for the extended persistence module XPH(M, f). In constructing these sets we use the fact that every essential class is born somewhere in the ordinary parameter range and then dies somewhere in the relative parameter range.

The following corollary follows from Proposition 4.11 and Lemma 4.13.

Corollary 4.14 *Let $(M, \partial M)$ be an n-dimensional manifold with boundary and let $f : M \to \mathbb{R}$ be a Morse function. Then*

$$\mathfrak{b}_k^{\mathrm{ord}}(M, f) \cup \mathfrak{d}_{k-1}^{\mathrm{ord}}(M, f) = \{(f(p), \mathrm{ord})| \ p \in \mathrm{Crit}(f, k) \cup \mathrm{Crit}(f, (k, +1))\}.$$

and

$$\mathfrak{b}_k^{\mathrm{rel}}(M, f) \cup \mathfrak{d}_{k-1}^{\mathrm{rel}}(M, f) = \{(f(p), \mathrm{rel})| \ p \in \mathrm{Crit}(f, n - k) \cup \mathrm{Crit}(f, (n - k - 1, -1))\}.$$

4.2 Relating the extended persistent homology of a manifold to that of its boundary

We can now restrict to the situation of interest for the XPHT; that of computing the extended persistent homology of a height function over a compact n-dimensional manifold with boundary embedded in $\mathbb{R}^n$. The results in this section start by comparing the sets of birth and death parameters for the height filtration of the manifold and for its boundary, in Propositions 4.15 and 4.16. The next step is to show that these births and deaths are paired consistently as endpoints of intervals in the relevant persistence modules (Theorem 4.17). We finish with a complete characterisation of the extended persistent homology for the manifold as a submodule of that for its boundary in Theorem 4.18.

The height function is specified in a direction v and restricted to various subsets of $\mathbb{R}^n$. That is, $h_v : \mathbb{R}^n \to \mathbb{R}$ with $h_v(x) = x \cdot v$. To ease notation let h_v^S denote the restriction of the height function to $S \subset \mathbb{R}^n$, that is $h_v^S = h_v|_S$.

Proposition 4.15 *Let $M \subset \mathbb{R}^n$ be a compact n-manifold with boundary. Suppose that $h_v^M : M \to \mathbb{R}$, the height function in direction v, is a Morse function. For each critical value t let $p(t)$ be the unique critical point of h_v^M or $h_v^{\partial M}$ with $h_v(p) = t$. For all $k > 0$ we have*

$$\mathfrak{b}_k^{\mathrm{ord}}(M, h_v^M) = \{(t, \mathrm{ord}) \in \mathfrak{b}_k^{\mathrm{ord}}(\partial M, h_v^{\partial M}) : \mathrm{sgn}(h_v^M, p(t)) = +1))\}$$
$$\mathfrak{b}_k^{\mathrm{rel}}(M, h_v^M) = \{(t, \mathrm{rel}) \in \mathfrak{b}_k^{\mathrm{rel}}(\partial M, h_v^{\partial M}) : \mathrm{sgn}(h_v^M, p(t)) = -1\}$$
$$\mathfrak{d}_k^{\mathrm{ord}}(M, h_v^M) = \{(t, \mathrm{ord}) \in \mathfrak{d}_k^{\mathrm{ord}}(\partial M, h_v^{\partial M}) : \mathrm{sgn}(h_v^M, p(t)) = +1\}$$
$$\mathfrak{d}_k^{\mathrm{rel}}(M, h_v^M) = \{(t, \mathrm{rel}) \in \mathfrak{d}_k^{\mathrm{rel}}(\partial M, h_v^{\partial M}) : \mathrm{sgn}(h_v^M, p(t)) = -1\}$$

Proof Choose $R > 0$ large enough so that $M \subset B(0, R)$ where $B(0, R)$ is the open ball of radius R centred on the origin. Let $L = \overline{B(0, R)} \setminus \mathrm{int}(M)$. As there are only finitely

many critical points of h_v^M and $h_v^{\partial M}$ there is an $\epsilon > 0$ such that all the critical values are at least ϵ apart. The critical values lie within $[\inf(h_v(M)), \sup(h_v(M))] \subset (-R, R)$ so we can reduce ϵ to be small enough that no critical value is within ϵ of $-R$ or R.

The function h_v defined over all of $\mathbb{R}^n$ has no critical points, so there will be no critical points in the interior of M. This means we need only consider critical points of $h_v^{\partial M}$.

For each $s \in \mathbb{R}$ we consider the sublevel sets of h_v restricted to the three subsets: M_s, $(\partial M)_s$ and L_s. By construction $M_s \cap L_s = (\partial M)_s$ and $M_s \cup L_s = h_v^{-1}(-\infty, s] \cap \overline{B(0, R)}$. For each $k > 0$ we therefore have $H_{k+1}(M_s \cup L_s) = 0 = H_k(M_s \cup L_s)$. Using this in the Mayer–Vietoris sequence shows us that $H_k((\partial M)_s)$ and $H_k(M_s) \oplus H_k(L_s)$ are isomorphic and hence

$$\beta_k(M_s) + \beta_k(L_s) = \beta_k((\partial M)_s) \tag{1}$$

for all $s \in \mathbb{R}$. For $k = 0$ we know $H_0(M_s \cup L_s) = 1$ whenever $s \geq -R$. Mayer–Vietoris then gives the short exact sequence

$$0 \to H_0((\partial M)_s) \to H_0(M_s) \oplus H_0(L_s) \to H_0(M_s \cup L_s) \to 0.$$

By comparing the ranks we have

$$\beta_0(M_s) + \beta_0(L_s) = \beta_0((\partial M)_s) + 1 \tag{2}$$

whenever $s > -R$.

Suppose that $(t, \mathrm{ord}) \in \mathfrak{b}_k^{\mathrm{ord}}(M, h_v)$ and thus $\beta_k(M_{t+\epsilon}) - \beta_k(M_{t-\epsilon}) = 1$. By Proposition 4.11 we know $\mathrm{sgn}(h_v^M, p(t)) = +1$ and this implies $\mathrm{sgn}(h_v^L, p(t)) = -1$. Proposition 4.11 now implies that $\beta_k(L_{t+\epsilon}) = \beta_k(L_{t-\epsilon})$. For $k > 0$ we can use (1) to calculate

$$\beta_k((\partial M)_{t+\epsilon}) - \beta_k((\partial M)_{t-\epsilon})$$
$$= (\beta_k(M_{t+\epsilon}) + \beta_k(L_{t+\epsilon})) - (\beta_k(M_{t-\epsilon}) + \beta_k(L_{t-\epsilon}))$$
$$= 1.$$

If $k = 0$ we instead use (2) to calculate

$$\beta_0((\partial M)_{t+\epsilon}) - \beta_0((\partial M)_{t-\epsilon})$$
$$= (\beta_0(M_{t+\epsilon}) + \beta_0(L_{t+\epsilon}) - 1) - (\beta_0(M_{t-\epsilon}) + \beta_0(L_{t-\epsilon}) - 1)$$
$$= 1.$$

This is where we use the requirement that ϵ is small enough that all critical points of M are greater than $-R + \epsilon$. Since $h_v^{\partial M}$ is Morse and t is the only critical value of $h_v^{\partial M}$ in $[t - \epsilon, t + \epsilon]$ we thus conclude that there is a birth event at t, that is $(t, \mathrm{ord}) \in \mathfrak{b}_k^{\mathrm{ord}}(\partial M, h_v)$.

Now suppose that $(t, \mathrm{ord}) \in \mathfrak{b}_k^{\mathrm{ord}}(\partial M, h_v)$ with $\mathrm{sgn}(h_v^M, p(t)) = +1$. This means $\beta_k((\partial M)_{t+\epsilon}) - \beta_k((\partial M)_{t-\epsilon}) = 1$, and $\mathrm{sgn}(h_v^L, p(t)) = -1$. Proposition 4.11 again tells us that $\beta_k(L_{t+\epsilon}) = \beta_k(L_{t-\epsilon})$ and using (1) we calculate

$$
\begin{aligned}
\beta_k(M_{t+\epsilon}) &- \beta_k(M_{t-\epsilon}) \\
&= (\beta_k((\partial M)_{t+\epsilon}) - \beta_k(L_{t+\epsilon})) - (\beta_k((\partial M)_{t-\epsilon}) - \beta_k(L_{t-\epsilon})) \\
&= 1.
\end{aligned}
$$

If $k = 0$ then we instead use (2) to calculate

$$
\begin{aligned}
\beta_0(M_{t+\epsilon}) &- \beta_0(M_{t-\epsilon}) \\
&= (\beta_k((\partial M)_{t+\epsilon}) - \beta_k(L_{t+\epsilon}) + 1) - (\beta_k((\partial M)_{t-\epsilon}) - \beta_k(L_{t-\epsilon}) + 1) \\
&= 1.
\end{aligned}
$$

We have again used $t - \epsilon > -R$. Since t is the only critical value of $h_v^{\partial A}$ in $[t - \epsilon, t]$ we conclude that $(t, \mathrm{ord}) \in \mathfrak{b}_k^{\mathrm{ord}}(M, h_v)$.

When considering the sets of births and deaths in the relative parameter range we need to use a relative version of the Mayer–Vietoris sequence. For this recall that $M \cap L = \partial M$, and $M^s \cap L^s = \partial M^s$. The relative version of the Mayer–Vietoris sequence states that there is a long exact sequence

$$
\begin{aligned}
\cdots &\to H_{k+1}(M \cup L, M^s \cup L^s) \to H_k((\partial M), (\partial M)^s) \to \cdots \\
&\to H_k(M, M^s) \oplus H_k(L, L^s) \to H_k(M \cup L, M^s \cup L^s) \to \cdots
\end{aligned}
$$

Since $M \cup L = B(0, R)$, and $M^s \cup L^s = B(0, R) \cap h_v^{-1}[s, \infty)$ we have (for $k \geq 0$ this time) $H_{k+1}(M \cup L, M^s \cup L^s) = 0 = H_k(M \cup L, M^s \cup L^s)$ for $s < R$. This implies $H_k((\partial M), (\partial M)^s)$ and $H_k(M, M^s) \oplus H_k(L, L^s)$ are isomorphic and hence

$$
\beta_k(\partial M, (\partial M)^s) = \beta_k(M, M^s) + \beta_k(L, L^s)
$$

for all $s < R$.

Suppose $(t, \mathrm{rel}) \in \mathfrak{b}_k^{\mathrm{rel}}(M, h_v^M)$ and thus $\beta_k(M, M^{t-\epsilon}) - \beta_k(M, M^{t+\epsilon}) = 1$. As $(t, \mathrm{rel}) \in \mathfrak{b}_k^{\mathrm{rel}}(M, h_v^M)$ we have $\mathrm{sgn}(h_v^M, p(t)) = -1$ we thus $\mathrm{sgn}(h_v^L, p(t)) = +1$.

From Corollary 4.13 we know that $\beta_k(L, L^{t-\epsilon}) = \beta_k(L, L^{t+\epsilon})$. We then calculate

$$
\begin{aligned}
\beta_k(\partial M, (\partial M)^{t-\epsilon}) &- \beta_k(\partial M, (\partial M)^{t+\epsilon}) \\
&= (\beta_k(M, M^{t-\epsilon}) + \beta_k(L, L^{t-\epsilon})) - (\beta_k(M, M^{t+\epsilon}) + \beta_k(L, L^{t+\epsilon})) \\
&= 1.
\end{aligned}
$$

Since t is the only critical value of $h_v^{\partial M}$ in $[t - \epsilon, t + \epsilon]$ we conclude that $(t, \mathrm{rel}) \in \mathfrak{b}_k^{\mathrm{rel}}(\partial M, h_v)$.

Now suppose that $(t, \text{rel}) \in \mathfrak{b}_k(\partial M, h_v)$ with $\text{sgn}(h_v^M, p(t)) = -1$ These facts imply $\beta_k(\partial M, (\partial M)^t) - \beta_k(\partial M, (\partial M)^{t+\epsilon}) = 1$ and $\text{sgn}(h_v^L, p(t)) = +1$. By Corollary 4.13 we therefore have $\beta_k(L, L^{t-\epsilon}) = \beta_k(L, L^{t+\epsilon})$ and can calculate

$$\beta_k(M, M^{t-\epsilon}) - \beta_k(M, M^{t+\epsilon})$$
$$= (\beta_k(\partial M, (\partial M)^{t-\epsilon}) - \beta_k(L, L^{t-\epsilon})) - (\beta_k(\partial M, (\partial M)^{t+\epsilon}) - \beta_k(L, L^{t+\epsilon}))$$
$$= 1.$$

Since t is the only critical value of $h_v^{\partial M}$ in $[t - \epsilon, t + \epsilon]$ we conclude that $(t, \text{rel}) \in \mathfrak{b}_k^{\text{rel}}(M, h_v)$.

The proof for the sets of death critical values is highly analogous; the difference of the Betti numbers is -1 instead of 1. $\qquad\square$

Throughout the following collection of results we fix the following sets: Let A be a compact subset of $\mathbb{R}^n$ whose boundary $\partial A = X$ is a finite collection of disjoint $n - 1$ manifolds. Let $R > 0$ such that $A \subset B(0, R)$. Let B be the set such that $A \cup B = \overline{B(0, R)}$ and $A \cap B = X$.

Let S_R denote the sphere of radius R. We can observe that $\partial A = X$ and $\partial B = X \cup S_R$. Let h_v be the height function in the direction $v \in S^{n-1}$, with v such that h_v^X is a Morse function.

Proposition 4.16 *Let $A \subset \mathbb{R}^n$ be a compact n-dimensional manifold with boundary. Let $h_v : \mathbb{R}^n \to \mathbb{R}$ be the height function in direction v such that h_v^X is a Morse function. Let $R > 0$ be such that $A \subset B(0, R)$. Let B be the set such that $A \cup B = \overline{B(0, R)}$ and $A \cap B = X$. Let S_R denote the sphere of radius R. Then we have the equality of the following disjoint unions:*

$$\mathfrak{b}_0(X, v) \sqcup \{(-R, \text{ord})\} = \mathfrak{b}_0(A, v) \sqcup \mathfrak{b}_0(B, v)$$
$$\mathfrak{b}_k(X, v) = \mathfrak{b}_k(A, v) \sqcup \mathfrak{b}_k(B, v) \qquad \text{for } k > 0$$
$$\mathfrak{d}_0(X, v) \sqcup \{(R, \text{ord})\} = \mathfrak{d}_0(A, v) \sqcup \mathfrak{d}_0(B, v)$$
$$\mathfrak{d}_k(X, v) = \mathfrak{d}_k(A, v) \sqcup \mathfrak{d}_k(B, v) \qquad \text{for } k > 0$$

Proof Since $\partial A = X$ we can use Proposition 4.15 to say

$$\mathfrak{b}_k(A, h_v) = \{(t, \text{ord}) \in \mathfrak{b}_k(X, h_v) | \; \text{sgn}(h_v^A, p(t)) = +1\}$$
$$\cup \{(t, \text{rel}) \in \mathfrak{b}_k(X, h_v) | \; \text{sgn}(h_v^A, p(t)) = -1\}$$
$$\mathfrak{d}_k(A, h_v) = \{(t, \text{ord}) \in \mathfrak{d}_k(X, h_v) | \; \text{sgn}(h_v^A, p(t)) = +1\}$$
$$\cup \{(t, \text{rel}) \in \mathfrak{d}_k(X, h_v) | \; \text{sgn}(h_v^A, p(t)) = -1\}.$$

Since $\partial B = X \sqcup S_R$ we can again apply Proposition 4.15 (now with B playing the role of M) to say

$$\mathfrak{b}_k(B, h_v) = \{(t, ord) \in \mathfrak{b}_k(X \sqcup S_R, h_v) | \operatorname{sgn}(h_v^B, p(t)) = +1\}$$
$$\cup \{(t, rel) \in \mathfrak{b}_k(X \sqcup S_R, h_v) | \operatorname{sgn}(h_v^B, p(t)) = -1\}$$
$$\mathfrak{d}_k(B, h_v) = \{(t, ord) \in \mathfrak{d}_k(X \sqcup S_R, h_v) \operatorname{sgn}(h_v^B, p(t)) = +1\}$$
$$\cup \{(t, rel) \in \mathfrak{d}_k(X \sqcup S_R, h_v) | \operatorname{sgn}(h_v^B, p(t)) = -1\}.$$

The critical points of h_v^B which lie on S_R are well understood. There are two critical points; one birth in ordinary homology degree 0 at $p_1 = -Rv$ with value $h_v(p_1) = -R$, and a death in relative homology degree 0 at $p_2 = Rv$ with $h_v(p_2) = R$. We thus can rewrite the birth and death sets of B as

$$\mathfrak{b}_0(B, h_v) = \{(-R, ord)\} \cup \{(t, ord) \in \mathfrak{b}_0(X, h_v) | \operatorname{sgn}(h_v^B, p(t)) = +1\}$$
$$\cup \{(t, rel) \in \mathfrak{b}_0(X, h_v) | \operatorname{sgn}(h_v^B, p(t)) = -1\}$$
$$\mathfrak{d}_0(B, h_v) = \{(R, rel)\} \cup \{(t, ord) \in \mathfrak{d}_0(X, h_v) \operatorname{sgn}(h_v^B, p(t)) = +1\}$$
$$\cup \{(t, rel) \in \mathfrak{d}_0(X, h_v) | \operatorname{sgn}(h_v^B, p(t)) = -1\}.$$

and for $k > 0$ we have

$$\mathfrak{b}_k(B, h_v) = \{(t, ord) \in \mathfrak{b}_k(X, h_v) | \operatorname{sgn}(h_v^B, p(t)) = +1\}$$
$$\cup \{(t, rel) \in \mathfrak{b}_k(X, h_v) | \operatorname{sgn}(h_v^B, p(t)) = -1\}$$
$$\mathfrak{d}_k(B, h_v) = \{(t, ord) \in \mathfrak{d}_k(X, h_v) \operatorname{sgn}(h_v^B, p(t)) = +1\}$$
$$\cup \{(t, rel) \in \mathfrak{d}_k(X, h_v) | \operatorname{sgn}(h_v^B, p(t)) = -1\}.$$

Since every critical point $p(t) \in X$ must be either $(+)$-critical or $(-)$-critical, by taking the union we get the statement of the proposition. □

Propositions 4.15 and 4.16 have shown how the sets of birth and death parameters for X, A, and B are related. The following theorem proves the much stronger result that the pairing of endpoints of the bars is consistent, and so we have isomorphisms between various extended persistence modules. This theorem is not a new result—it was proved using Alexander duality in Edelsbrunner and Kerber (2012). We believe our Morse-theoretic proof may be more readily adapted to other scenarios.

Theorem 4.17 *Let A, B, and X be as in Proposition 4.16. We have*

$$\operatorname{XPH}_k(X, v) \oplus \operatorname{XPH}_k(A \cup B, v) = \operatorname{XPH}_k(A, v) \oplus \operatorname{XPH}_k(B, v).$$

That is $\operatorname{XPH}_0(X, v)) \oplus \mathcal{I}_{((-R, ord), (R, rel))} = \operatorname{XPH}_0(A, v) \oplus \operatorname{XPH}_0(B, v)$ *and for* $k > 0$ *we have* $\operatorname{XPH}_k(X, v)) = \operatorname{XPH}_k(A, v) \oplus \operatorname{XPH}_k(B, v).$

Proof Let us first consider the case where $k > 0$. Since $X \subset A$ and $X \subset B$ we have an induced morphisms on persistence modules

$$\varphi : \operatorname{XPH}_k(X, v) \rightarrow \operatorname{XPH}_k(A, v) \oplus \operatorname{XPH}_k(B, v).$$

Furthermore from the ordinary and relative versions of the Mayer–Vietoris sequence we know $\varphi_{(t,ord)}$, and $\varphi_{(t,rel)}$ are both isomorphisms for all $t \in \mathbb{R}$. This implies that φ is must be injective.

Injective morphisms between persistence modules were studied extensively in Bauer et al. (2014). Bauer and Lesnick showed that an injective morphism will induce an injective map ρ on the sets of intervals in the interval decomposition of $\mathrm{XPH}_k(X, v)$ to those in the interval decomposition of $\mathrm{XPH}_k(A, v) \oplus \mathrm{XPH}_k(B, v)$ such that every interval in $[b, d)$ in $\mathrm{XPH}_k(X, v)$ is mapped to an interval $\rho([b, d)) = [b', d)$ with the same death time and $b' \leq b$.

By Proposition 4.16 we know that the sets of start and end parameters for the barcode decompositions satisfy $\mathfrak{b}_k(X, v) = \mathfrak{b}_k(A, v) \cup \mathfrak{b}_k(B, v)$. As the two persistence modules have the same number of intervals, the matching ρ must in fact be a bijection. Observe that if $f : S \to S$ is a bijection from a finite set to itself such that $f(s) \leq s$ for all $s \in S$ then we are forced to have f the identity. This argument shows $\rho([b, d)) = \rho([b, d))$ and the interval decompositions of $\mathrm{XPH}_k(X, v)$ and $\mathrm{XPH}_k(A, v) \oplus \mathrm{XPH}_k(B, v)$ are the same and they are isomorphic as persistence modules.

For the case where $k = 0$ we need to consider the complication of the homology class corresponding to the sphere S_R. We know from Proposition 4.16 that $\mathfrak{b}_0(X, v) \sqcup \{(-R, ord)\} = \mathfrak{b}_0(A, v) \sqcup \mathfrak{b}_0(B, v)$, which we will denote $\mathfrak{b}$, and $\mathfrak{d}_0(X, v) \sqcup \{(R, ord)\} = \mathfrak{d}_0(A, v) \sqcup \mathfrak{d}_0(B, v)$, which we will denote $\mathfrak{d}$. This means that we can define a bijection $\rho : \mathfrak{b} \to \mathfrak{b}$ such that $\rho(b) = b'$ if there exists a d such that

$$(b, d] \in \mathrm{XPH}_0(X, v)) \oplus \mathcal{I}_{((-R, ord),(R, rel))}$$

and

$$(b', d] \in \mathrm{XPH}_k(A, v) \oplus \mathrm{XPH}_k(B, v).$$

Observe that $[(-R, ord), (R, rel))$ is an interval in the interval decomposition of $\mathrm{XPH}_0(B, v)$, it corresponds to the connected component containing S_R. This implies that $\rho((-R, ord)) = (-R, ord)$. Just as in the case for $k > 0$ we can consider the ordinary and relative Mayer Vietoris sequences to show that the morphisms $H_0(X_t) \to H_0(A_t) \oplus H_0(B_t)$ and $H_0(X, X^t) \to H_0(A, A^t) \oplus H_0(B, B^t)$ induced by inclusions are injective for all t and hence the morphism $\varphi : \mathrm{XPH}_0(X, v) \to \mathrm{XPH}_0(A, v) \oplus \mathrm{XPH}_0(B, v)$ is injective. Again this implies there is an injective map which pairs each interval $[b, d)$ in $\mathrm{XPH}_0(X, v)$ to an interval $[b', d)$ with the same death time and $b' \leq b$. This implies that our function $\rho : \mathfrak{b} \to \mathfrak{b}$ has $\rho(b) \leq b$ for all $b \in \mathfrak{b}_0(\mathrm{XPH}_0(X))$. Together these imply $\rho(b) \leq b$ for all $b \in \mathfrak{b}$ which, since $\mathfrak{b}$ is finite, implies ρ is the identity. Hence the interval decompositions of $\mathrm{XPH}_0(X, v) \oplus \mathcal{I}_{[(-R, ord),(R, rel))}$ and $\mathrm{XPH}_0(A, v) \oplus \mathrm{XPH}_0(B, v)$ are the same and they are isomorphic as persistence modules. $\square$

Combining Theorem 4.17 with Proposition 4.16 allows us to express the extended persistent homology of a height function over A as a nice submodule of the extended persistent homology of that same height over ∂A.
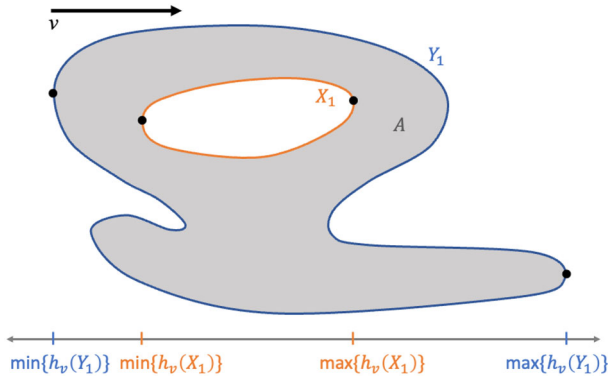
Fig. 3 *A* is a connected subset of the plane with one interior boundary component X_1 in orange, and one exterior boundary component Y_1 in blue. Maximum and minimum values and critical points for the height function h_v are marked for both boundary components (color figure online)

Theorem 4.18 *Let $A \subset \mathbb{R}^n$ be an n-manifold with boundary $X = \partial A$. Let v be a direction such that $h_v^A : A \to \mathbb{R}$ is a Morse function. Let the interval decomposition of the degree-k extended persistent homology of $h_v^X : X \to \mathbb{R}$ be*

$$\mathrm{XPH}_k(X, h_v) = \bigoplus_{[b_i, d_i) \in S_X} \mathcal{I}_{[b_i, d_i)}.$$

Let J_A^k be the subset of intervals $[b_i, d_i)$ such that $b_i = (h_v(p), \mathrm{ord})$ for some $p \in \mathrm{Crit}(h_v^A, (k, +1))$, or $b_i = (h_v(p), \mathrm{rel})$ for some $p \in \mathrm{Crit}(h_v^A, (n - k - 1, -1))$. Then

$$\mathrm{XPH}_k(A, h_v) = \bigoplus_{[b_i, d_i) \in J_A^k} \mathcal{I}_{[b_i, d_i)}.$$

We can more readily describe the essential classes in degrees 0 and $n - 1$ in terms of the minimum and maximum values on the different connected components of the boundary. Observe that a compact connected $(n - 1)$-dimensional manifold Y embedded in $\mathbb{R}^n$ separates $\mathbb{R}^n$ into two connected open sets, one of which is 'inside' Y and one of which is 'outside' (this is the unbounded component of the two). This theorem is known as the Jordan-Brouwer separation theorem. We use this to define the connected components of $X = \partial A$ as interior or exterior boundary components.

Definition 4.19 Let $A \subset \mathbb{R}^n$ be a compact n-dimensional manifold with boundary $\partial A = X$. Let $\hat{X}$ be a connected component of X, and $\hat{A}$ the connected component of A that contains $\hat{X}$. We say that $\hat{X}$ is an *interior boundary component* if $\hat{A} \backslash \hat{X}$ is contained in the unbounded connected component of $\mathbb{R}^n \backslash \hat{X}$. We say that $\hat{X}$ is an *exterior boundary component* if $\hat{A} \backslash \hat{X}$ is contained in the bounded connected component of $\mathbb{R}^n \backslash \hat{X}$.

See Fig. 3 for an illustration of the above definition and the following proposition.

Proposition 4.20 *Let $A \subset \mathbb{R}^n$ be an n-manifold with boundary $X = \partial A$. Let v be a direction such that $h_v^A : A \to \mathbb{R}$ is a Morse function. Let $\{X_1, \ldots X_k\}$ be the interior boundary components of X and $\{Y_1, \ldots Y_l\}$ be the exterior boundary components of X. Then*

$$\mathrm{Ess}_0(A, h_v) = \sum_{j=1}^{l} \mathcal{I}_{[(\min\{h_v(Y_j)\}, \mathrm{ord}), (\max\{h_v(Y_j)\}, \mathrm{rel}))}$$

and

$$\mathrm{Ess}_{n-1}(A, h_v) = \sum_{i=1}^{k} \mathcal{I}_{[(\max\{h_v(X_i)\}, \mathrm{ord}), (\min\{h_v(X_i)\}, \mathrm{rel}))}.$$

Proof If A is the disjoint union of $A_1, \ldots A_l$ then $\mathrm{XPH}(A) = \oplus_{i=1}^{l} \mathrm{XPH}(A_i)$. This means that it is sufficient to prove the case where A is connected. We assume A is connected for the remainder of the proof.

Observe that for M a connected $(n-1)$-dimensional manifold we have $\beta_0(M) = 1 = \beta_{n-1}(M)$ so there is exactly one essential persistent homology interval module in each of these homology degrees for extended persistent homology of M with respect to h_v.

The interval in $\mathrm{Ess}_0(M, h_v)$ is born at the first appearance of M, that is at $(\min\{h_v(M)\}, \mathrm{ord})$. Since M is connected we have this homology class is trivial in $H_0(M, L)$ for any non-empty subset $L \subset M$. This implies that the death of this interval in $\mathrm{Ess}_0(M, h_v)$ is at parameter $(\max\{h_v(M)\}, \mathrm{rel})$. We have shown that

$$\mathrm{Ess}_0(M, h_v) = \mathcal{I}_{[(\min\{h_v(M)\}, \mathrm{ord}), (\max\{h_v(M)\}, \mathrm{rel}))}.$$

Using the symmetry of extended persistent homology for manifolds (Proposition 2.8) we have

$$\mathrm{Ess}_{n-1}(M, h_v) = \mathcal{I}_{[(\max\{h_v(M)\}, \mathrm{ord}), (\min\{h_v(M)\}, \mathrm{rel}))}.$$

Since X is the disjoint union of the interior boundary components $\{X_1, \ldots X_k\}$ and the exterior boundary component Y we have

$$\mathrm{Ess}_0(X, h_v) = \left(\oplus_{i=1}^{n} \mathcal{I}_{[(\min\{h_v(X_i)\}, \mathrm{ord}), (\max\{h_v(X_i)\}, \mathrm{rel}))} \right)$$
$$\oplus \mathcal{I}_{[(\min\{h_v(Y)\}, \mathrm{ord}), (\max\{h_v(Y)\}, \mathrm{rel}))}$$

and

$$\mathrm{Ess}_{n-1}(X, h_v) = \left(\oplus_{i=1}^{n} \mathcal{I}_{[(\max\{h_v(X_i)\}, \mathrm{ord}), (\min\{h_v(X_i)\}, \mathrm{rel}))} \right)$$
$$\oplus \mathcal{I}_{[(\max\{h_v(Y)\}, \mathrm{ord}), (\min\{h_v(Y_j)\}, \mathrm{rel}))} \Big).$$

We can use Theorem 4.18 to deduce $\mathrm{Ess}_0(A, h_v)$ and $\mathrm{Ess}_{n-1}(A, h_v)$ from the various persistence modules $\mathrm{Ess}_0(X_i, h_v), \mathrm{Ess}_{n-1}(X_i, h_v), \mathrm{Ess}_0(Y, h_v)$ and $\mathrm{Ess}_{n-1}(Y, h_v)$.

Consider an interior boundary component X_i, and let $p_i \in X_i$ be the global minimum of $h_v^{X_i}$. We know that A is contained in the infinite component of $\mathbb{R}^n \backslash X_i$, and p_i must be a $(-)$-critical point for h_v^A. This implies that $[(\min\{h_v(X_i)\}, \mathrm{ord}), (\max\{h_v(X_i)\}, \mathrm{rel}))$ is not included in J_A^0 (where J_A^k is defined in the statement of Theorem 4.18). Similarly let $\hat{x}_i$ denote the global maximum of h_v over X_i. We know that A is contained in the infinite component of $\mathbb{R}^n \backslash X_i$, and $\hat{p}_i$ must be a $(+)$-critical point for h_v^A. This implies that

$$[(\max\{h_v(X_i)\}, \mathrm{ord}), (\min\{h_v(X_i)\}, \mathrm{rel})) \in J_A^{n-1}.$$

If we instead consider the exterior boundary component Y then the global minimum of h_v^Y will be a $(+)$-critical point for h_v^A and the global maximum of h_v^Y will be a $(-)$-critical point for h_v^A. This implies that

$$[(\min\{h_v(Y)\}, \mathrm{ord}), (\max\{h_v(Y)\}, \mathrm{rel})) \in J_A^0$$

but we do not include $[(\max\{h_v(Y)\}, \mathrm{ord}), (\min\{h_v(Y)\}, \mathrm{ord}))$ in J_A^{n-1}. □

5 The extended persistent homology transform

5.1 Background

The persistent homology transform (PHT) maps the space of shapes embedded in Euclidean space into a space of topological summaries. Instead of comparing the original shapes we can compare their topological transforms. The philosophy is that the persistent homology of a height function in some direction v records geometric information from the perspective of direction v. As v changes, the persistent homology classes track geometric features in M. The key insight behind the persistent homology transform (PHT) is that by considering the persistent homology from every direction, we preserve all information about the shape.

Before giving the formal definition we should first identify the subsets of space which are allowable shapes, that is the domain of the PHT. We will want our subsets to be reasonably nice. The most general setting for which theoretical properties about the PHT are proved are compact o-minimal sets, which are called *constructible* in Curry et al. (2022). For the purposes of this paper it is sufficient to know that compact and semi-algebraic or piecewise linear are sufficient conditions for a subset of Euclidean space to be constructible. We will denote the space of constructible subsets of $\mathbb{R}^n$ by $\mathrm{CS}(\mathbb{R}^n)$.

Given an constructible set $M \subset \mathbb{R}^n$, and $v \in S^{n-1}$, let h_v be the corresponding to a height function in direction v,

$$h_v : M \to \mathbb{R}$$

$$h_v : x \mapsto \langle x, v \rangle.$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. We can construct a persistence module $\mathrm{PH}_k(M, h_v)$ by filtering M by the sub-level sets of h_v and taking degree-k homology groups. The underlying parameter set for the persistence module is $\mathbb{R}$, the attached vector space at $t \in \mathbb{R}$ is $H_k(h_v^{-1}(-\infty, t])$, and for $s \leq t$ the transition map φ_s^t is the induced map on homology from the inclusion $h_v^{-1}(-\infty, s] \subseteq h_v^{-1}(-\infty, t]$.

Let $\mathrm{PM}(\mathbb{R})$ denote the standard space of persistence modules over parameter space $\mathbb{R}$.

Definition 5.1 The *Persistent Homology Transform* PHT of a constructible set $M \in \mathrm{CS}(\mathbb{R}^n)$ is the map $\mathrm{PHT}(M) : S^{n-1} \to \mathrm{PM}(\mathbb{R})^n$ that sends a direction to the set of persistence modules by filtering M in the direction of v:

$$\mathrm{PHT}(M) : v \mapsto \left(\mathrm{PH}_0(h_v^M), \mathrm{PH}_1(h_v^M), \ldots, \mathrm{PH}_{n-1}(h_v^M) \right)$$

where $h_v : M \to \mathbb{R}$, $h_v(x) = \langle x, v \rangle$ is the height function on M in direction v.

Various properties of the PHT have been proved in Turner et al. (2014), Ghrist et al. (2018), Curry et al. (2022). Stability results bound the distance between h_v and h_w when $v, w \in S^{n-1}$ are close. This implies that for each $M \in \mathrm{CS}(R^n)$, its persistent homology transform, $\mathrm{PHT}(M)$, is a continuous function over S^{n-1} when we equip PM with a Wasserstein metric.

Another very important property about the PHT is its injectivity, that is that for $M_1, M_2 \subset \mathbb{R}^n$, if $\mathrm{PHT}(M_1) = \mathrm{PHT}(M_2)$ then $M_1 = M_2$ as subsets of $\mathbb{R}^n$. This was originally proved in Turner et al. (2014) for piecewise linear compact subsets in $\mathbb{R}^2$ and $\mathbb{R}^3$, and then the more general proof was given in Curry et al. (2022) and independently in Ghrist et al. (2018).

We can now define a distance between M_1, M_2 constructible sets via their persistent homology transforms. We basically just integrate the Wasserstein distances over all the possible directions.

Definition 5.2 Fix $p \in [1, \infty)$ and ambient dimension n. Define the distance function $d_p^{\mathrm{PHT}} : \mathrm{CS}(\mathbb{R}^n) \times \mathrm{CS}(\mathbb{R}^n) \to \mathbb{R}$ by

$$d_p^{\mathrm{PHT}}(M_1, M_2)^p = \int_{v \in S^{n-1}} \sum_{k=0}^{n-1} W_p(\mathrm{PH}_k(M_1, h_v), \mathrm{PH}_k(M_2, h_v))^p \, dv.$$

5.2 Now with extended persistence

We can define a new distance function over $\mathrm{CS}(\mathbb{R}^n)$ by replacing the normal persistent homology with extended persistent homology. We can construct a definition of a distance between extended persistent homology transforms by replacing the Wasserstein distance between the original persistence modules with those between extended persistence modules.

Definition 5.3 Fix $p \in [1, \infty)$ and ambient dimension n. Define the distance function $d_p^{\text{XPHT}} : \text{CS}(\mathbb{R}^n) \times \text{CS}(\mathbb{R}^n) \to \mathbb{R}$ by

$$d_p^{\text{XPHT}}(M_1, M_2)^p = \int_{v \in S^{n-1}} \sum_{k=0}^{n-1} W_p(\text{XPH}_k(M_1, h_v), \text{XPH}_k(M_2, h_v))^p \, dv.$$

For the PHT one theoretical result was the continuity of the PHT(M) as a function from S^{n-1}. This continuity justified the approximation of the PHT by a finite subset of directions. The proofs for the continuity of the PHT can be easily modified to show continuity of the XPHT. Let $\mathscr{E}$ denote the space of extended persistence modules. Then for all $M \in \text{CS}(\mathbb{R}^n)$, the function $\text{XPHT}_k(M) : S^{n-1} \to \mathscr{E}$ is continuous when we equip $\mathscr{E}$ with the p-Wasserstein distance (for $p \in [1, \infty)$), or the bottleneck distance.

In Skraba and Turner (2020) a stability result for the PHT was proven in the case where M_1 and M_2 were different embeddings of the same simplicial complex. This bounded the distance between PHT(M_1) and PHT(M_2) in terms of the distances between the sets of vertices in the embedding. The proof of this stability theorem can be easily modified to prove an analogous statement for the extended persistent homology transform.

Since the extended persistence module for a filtration by a height function contains strictly more information than the regular persistence module for that height function, the injectivity results for the PHT will automatically also hold for the XPHT.

6 Application to binary images

In this section we describe how to interpret a binary digital image as a PL-manifold with boundary, construct boundary curves as PL 1-manifolds, and adapt the results of Sect. 4 to this setting using a simulation of simplicity methodology (Edelsbrunner and Mücke 1990).

6.1 Boundary curves

A binary digital image is a two-dimensional array, P, with elements called *pixels* taking values in $\{0, 1\}$. The array is indexed by integers $1 \le i \le m$ and $1 \le j \le n$, so that $P(i, j)$ is the element in the ith row and jth column of P. We can also treat pixels as points in the plane by mapping the array index to a Cartesian coordinate (the first axis is oriented down the page and second from left to right). Those pixels taking the value '1' are defined to be the *foreground* $F := P^{-1}[1]$ and those with value '0' are the *background* $G := P^{-1}[0]$. A small patch of such a binary image array is illustrated in Fig. 4.

To answer questions about the connectivity of objects represented by the image, we must define a neighbourhood or adjacency relation for each pixel. Two standard options called 4- and 8-connectivity in digital topology are defined as follows.

Fig. 4 The rows and columns of a binary digital image are indexed by i and j respectively. Foreground pixels are labelled '1' and connected when 8-adjacent. Background pixels are labelled '0' and connected when 4-adjacent. Segments of the boundary curves are drawn in orange and the boundary point labelled 'p' has coordinates $(i + \frac{1}{2}, j)$ (color figure online)
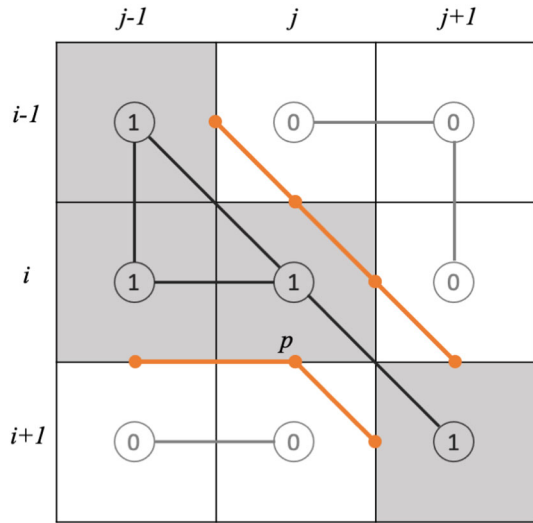

Definition 6.1 A pixel (k, l) is said to be a *4-adjacent* or *direct* neighbour of (i, j) if their ℓ_1 distance is exactly 1: $|i - k| + |j - k| = 1$. Pixels are *8-adjacent* neighbours if the ℓ_∞ distance is 1: $\max\{|i - k|, |j - k|\} = 1$. The 4-neighbourhood of pixel (i, j) consists of its four 4-adjacent neighbours and the 8-neighbourhood is defined similarly.

The connectivity of a set of pixels is then determined according to a specified adjacency relation. If we choose to use the 8-neighbourhood for pixels in both the foreground and the background, however, counter-intuitive situations may arise such as a simple closed digital curve that does *not* separate the plane into two pieces. The resolution of this within digital topology is to treat pixels in the foreground as connected with respect to the 8-neighbourhood and pixels in the background with the 4-neighbourhood, or vice-versa (Kong and Rosenfeld 1989).

We now proceed to construct a set, $\mathcal{C}$, of piecewise-linear curves that subdivide the plane so that each connected component of $\mathbb{R}^2 \setminus \mathcal{C}$ contains pixels of only one type (either foreground or background), and such that the digital connected components of F and G are in one-to-one correspondence with those of $\mathbb{R}^2 \setminus \mathcal{C}$. As described above, we use 8-connectivity for the foreground and 4-connectivity for the background. We assume (and in practice add) a layer of background pixels to any given rectangular array, P, to ensure there is a single connected background component surrounding all other components.

Definition 6.2 *Boundary points*. For every pair of 4-adjacent pixels such that $P(i, j) = 1$ and $P(k, l) = 0$, define the boundary point $p = (\frac{1}{2}(i + k), \frac{1}{2}(j + l))$.

There are only four possible configurations. For example if $(i, j) \in F$ and its direct neighbour $(i + 1, j) \in G$, then $p = (i + \frac{1}{2}, j)$; the other three cases are simple adjustments to this pattern. Note that since (i, j) and (k, l) are 4-adjacent, the boundary point has only one coordinate with the $\frac{1}{2}$ offset and one remaining an integer. See Fig. 4 for an illustrative example.
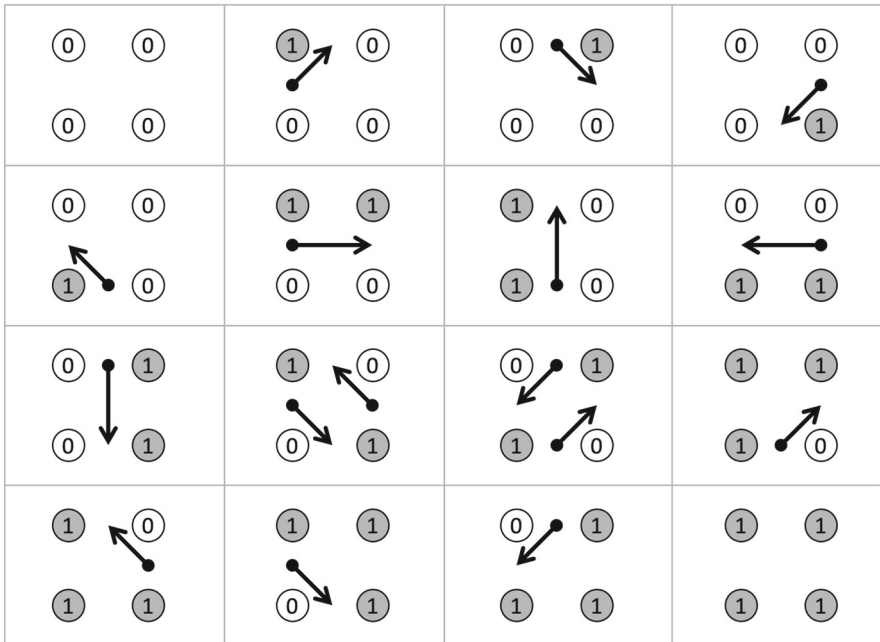
Fig. 5 Each of the 2^4 possible 2×2 binary-valued pixel patches showing the associated oriented boundary edges for the case that foreground pixels connect when 8-adjacent. The edge orientation always has the foreground on the left

The next step is to connect pairs of boundary points by line segments in such a way that the foreground and background pixel connectivities are respected. This is achieved by exhaustive enumeration of 2×2 pixel patches as illustrated in Fig. 5.

Lemma 6.3 *Let $\mathcal{C} \subset \mathbb{R}^2$ be the union of boundary points and edges derived from a binary digital array P. The set $\mathcal{C}$ is a disjoint union of simple closed piecewise linear curves.*

Proof Let P be the $n_r \times n_c$ array with rows indexed by $i = 1, \ldots, n_r$ and columns by $j = 1, \ldots, n_c$. By assumption, the outermost rows and columns of P are background, i.e., $P(1, j) = P(n_r, j) = P(i, 1) = P(i, n_c) = 0$. Each boundary point sits half-way between two 4-adjacent pixels with distinct values, so every boundary point has first coordinate $1 < p_1 < n_r$ and second coordinate $1 < p_2 < n_c$. It follows that every boundary point must belong to exactly two adjacent 2×2 pixel patches and that every boundary point connects to exactly two boundary edges, by construction. As a combinatorial object then, each component of $\mathcal{C}$ is a discrete closed 1-manifold. Also by construction (see Fig. 5) any two boundary edges can only intersect at their endpoints and we conclude that each component of $\mathcal{C}$ is a simple closed PL-curve. □

Lemma 6.4 *Let A be the union of components of $\mathbb{R}^2 \setminus \mathcal{C}$ that contain at least one foreground pixel of the binary image array P. Then A is a bounded manifold with boundary $\partial A = \mathcal{C}$.*

Proof A is bounded because the image array is finite. Each connected component $C_a \in \mathcal{C}$ is a simple closed PL-curve, so $\mathbb{R}^2 \setminus C_a$ consists of two open domains. Each connected component of $\mathbb{R}^2 \setminus \mathcal{C}$ is formed by the intersection of a finite number of these domains so is also open and it follows that A is open. Clearly $\partial A \subset \mathcal{C}$, we must now show that $\partial A \supset \mathcal{C}$. Let $p \in \mathcal{C}$, i.e., p is an arbitrary point on one of the boundary edge segments. We can write the coordinates of p as $(i + \epsilon, j + \eta)$ for integers i, j and fractional parts $\epsilon, \eta \in [0, 1)$. We know that each boundary edge divides the 2×2 patch with corners $(i, j), (i + 1, j), (i + 1, j + 1), (i, j + 1)$, into two pieces such that at least one of these corners has $P(k, l) = 1$ and this implies that $p \in \partial A$. □

The above results show that $\mathrm{cl}(A)$ and $X = \partial A$ satisfy the conditions for the theorem(s) of Sect. 4.2 as X is a finite union of disjoint piecewise-linear 1-manifolds. We then define B to be the closed complement of A in the rectangular domain of the image, $B = ([1, n_r] \times [1, n_c]) \setminus A$. A straightforward argument by contradiction shows that no background pixel lies in A, so we have $P^{-1}[0] \subset B$.

Remark 6.5 Given a three-dimensional binary array of voxels, $V(i, j, k)$, there are analogous definitions of direct-adjacency between elements, and results that require foreground and background to be viewed with complementary adjacencies to maintain topological consistency (Kong and Rosenfeld 1989). There are also established methods to construct a triangular mesh surface that separates the connected components of foreground and background. These are termed 'marching cubes algorithms' (Newman and Yi 2006).

6.2 Breaking ties and other practical considerations

In this section we derive additional results required to extend theorems from Sect. 4 so that they hold for the digital boundary curves. In particular, Theorem 4.18 specified that the height function in direction v is a Morse function, i.e., that the critical points are isolated and the critical values are distinct. Both these conditions are challenged by the geometry of a digital grid as the boundary curve points lie at integer and half-integer coordinates, and the boundary curve edges are either horizontal, vertical or in one of two diagonal directions. Additionally, the direction vectors v are typically chosen to be equal-spaced rational fractions of π, and will often be perpendicular to some boundary edges. This means that when computing the XPHT for equiangular directions v we expect many vertices of the boundary curves to have the same height with respect to any given v.

Our computations of persistent homology involve height filtrations of boundary curves considered as simplicial complexes. The algorithm for computing persistent homology of simplicial complexes orders simplices by their maximal vertex value with lower-dimensional simplices added before higher-dimensional ones if their maximal values are the same. It is well understood that the persistent homology of this discrete filtration of complexes gives the same persistence diagram as that of a continuous filtration of a PL-embedding of the complex. We do, however, need to explore how a filtration with multiple simplices taking the same height with respect to direction v

relates to the critical points of a piecewise-linear Morse function constructed from an arbitrarily close direction v_t.

We first need to generalise the notion of 0-critical point to allow for line segments

Definition 6.6 Let γ be a piecewise-linear simple closed curve in $\mathbb{R}^2$ with m vertices traversed in cyclic order, $x_0, x_1, \ldots, x_m = x_0$. Note that in the following text the indices are assumed to be given as integers modulo m. We say x_i is an *isolated* 0-*critical vertex* if $h_v(x_{i-1}) > h_v(x_i)$ and $h_v(x_i) < h_v(x_{i+1})$. We say that the line segment from x_j to x_k is a 0-*critical segment* if $h_v(x_i) = h_v(x_j)$ for all $i = j, j+1, \ldots, k$ and that $h_v(x_{j-1}) > h_v(x_j)$ and $h_v(x_k) < h_v(x_{k+1})$. Denote this line segment as $e(x_j, x_k)$.

Observe that if $e = e(x_j, x_k)$ is a 0-critical segment for h_v then the vector $x_k - x_j$ must be perpendicular to v, and h_v is constant over e. Recall that 0-critical points on the boundary correspond to local minima, and the 0-critical points which are (+)-critical will be local minima as points in A. To go from 0-critical points to 0-critical segments we need to relax this notion of minima to have non-strict inequalities.

Definition 6.7 We say that a vertex x_i lying on a 0-critical segment e is (+)-*critical* for h_v^A if there exists an $\epsilon > 0$ such that for all $a \in B(x_i, \epsilon) \cap A$ we have $h_v(x_i) \leq h_v(a)$. Given the definition of manifold with boundary, if any vertex on a 0-critical segment e is (+)-critical then every vertex on it will be, and we say that the 0-critical segment is (+)-*critical*.

We now distinguish which of the 0-critical points and segments on a piecewise linear boundary curve are (+)-critical. We will use the fact that the orientation of planar triangles is defined by the sign of the determinant of a matrix formed from edge vectors as follows. First let $DET(x, y)$ be the determinant of a 2×2 matrix with columns x and y. Given a triangle $\Delta(a, b, c)$ with positive area, the vertices a, b, c are in an anticlockwise order if $DET(c - b, a - b) > 0$ and in a clockwise order if $DET(c - b, a - b) < 0$.

The following two geometric lemmas cover the cases where one or both of the edges adjacent to a local minimum is perpendicular to the direction v.

Lemma 6.8 *Let $A \subset \mathbb{R}^2$ be a bounded subset whose boundary is the disjoint union of piecewise linear closed curves. Let $\gamma = (x_0, x_1, x_2, \ldots x_m = x_0)$ be a piecewise linear boundary curve of A with vertices listed anticlockwise with respect to A. Fix $v \in S^1$. If x_i is an isolated 0-critical vertex of h_v^γ, or an endpoint of a 0-critical segment e, then x_i is (+)-critical for h_v^A if and only if $DET(x_{i+1} - x_i, x_{i-1} - x_i) > 0$.*

Proof There is some $\epsilon > 0$ such that the interior of the triangle bounded by x_i, $x_i + \epsilon(x_{i-1} - x_i)$ and $x_i + \epsilon(x_{i+1} - x_i)$ is either entirely contained in A or is entirely contained in the complement of A. For the sake of computations let $y_{i-1} = x_i + \epsilon(x_{i-1} - x_i)$ and $y_{i+1} = x_i + \epsilon(x_{i+1} - x_i)$. By assumption we have $h_v(y_{i+1}) \geq h_v(x_i)$ and $h_v(y_{i-1}) \geq h_v(x_i)$ with at least one of these inequalities strict, which implies that the convex hull of x_i, y_{i-1} and y_{i+1} has positive area and $DET(y_{i+1} - x_i, y_{i-1} - x_i) \neq 0$.

Suppose that x_i is (+)-critical for h_v^A which implies that $\Delta(y_{i-1}, x_i, y_{i+1})$ is a subset of A. Since γ traces a boundary curve that is going anticlockwise around A we

must have vertices y_{i-1}, x_i, y_{i+1} in an anticlockwise order. This implies $DET(y_{i+1} - x_i, y_{i-1} - x_i) > 0$.

If x_i is not (+)-critical then the opposite holds: we have $\Delta(y_{i-1}, x_i, y_{i+1})$ is not contained in A, that y_{i-1}, x_i, y_{i+1} are in a clockwise order and thus $DET(y_{i+1} - x_i, y_{i-1} - x_i) < 0$. □

If the point x_i is contained strictly inside a 0-critical segment then we need an alternative approach. This will also be useful when successive points are close to co-linear with respect to a direction v, because we want to avoid any possible issues with floating point errors in computations.

Lemma 6.9 *Let $A \subset \mathbb{R}^2$ be a bounded subset whose boundary is the disjoint union of piecewise linear closed curves. Let $\gamma = (x_0, x_1, x_2, \ldots x_m = x_0)$ be a piecewise linear boundary curve of A with vertices listed anticlockwise with respect to A. Fix $v \in S^1$. Let x_i be an isolated 0-critical vertex or a vertex in a 0-critical segment e with respect to the function h_v^γ. Furthermore, suppose that $\Delta(x_{i-1}, x_i, x_{i+1})$ has an obtuse angle α_i at x_i. Let w_i denote the rotation of the vector $x_{i+1} - x_i$ anticlockwise by $\pi/2$. Then x_i is (+)-critical for h_v^A if and only if $w_i \cdot v > 0$.*

Proof If $w_i \cdot v = 0$ then we deduce that $h_v(x_i)$ lies strictly between $h_v(x_{i-1})$ and $h_v(x_{i+1})$ contradicting the assumption x_i is 0-critical, so we know $w_i \cdot v \neq 0$.

Since γ is traced anticlockwise around A and $\pi/2 < \alpha_i \leq \pi$ we know that w_i, the rotation of $x_{i+1} - x_i$ anticlockwise by $\pi/2$, will point into A from x_i, and if we rotate anticlockwise around x_i from the direction w_i we encounter $x_{i-1} - x_i$ at an angle strictly less than π. Set $y = x_i + w_i$. Since w_i points into A from x_i, for small $\epsilon > 0$, we can cover $A \cap B(x_i, \epsilon)$ by triangles $\Delta(x_{i-1}, x_i, y)$ and $\Delta(y, x_i, x_{i+1})$.

If $w_i \cdot v > 0$ then $h_v(y) > h_v(x_i)$. Every point $a \in \Delta(x_{i-1}, x_i, y)$ can be written as a convex combination $a = a_1 x_{i-1} + a_2 x_i + a_3 y$. For this a,

$$h_v(a) = a_1 h_v(x_{i-1}) + a_2 h_v(x_i) + a_3 h_v(y) \geq h_v(x_i)$$

as $h_v(x_{i-1}), h_v(y) \geq h_v(x_i)$. Similarly every point $a \in \Delta(y, x_i, x_{i+1})$ also satisfies $h_v(a) \geq h_v(x_i)$. Together these imply that x_i is (+)-critical.

If $w_i \cdot v < 0$ then h_v decreases along $t \mapsto x_i + t w_i$, showing that for all $\epsilon > 0$ there are points in $a_\epsilon \in A \cap B(x_i, \epsilon)$ with $h_v(a_\epsilon) < h_v(x_i)$. This implies x_i is not (+)-critical. □

We are now ready to generalise Theorem 4.18 to PL subsets of the plane where we drop the Morse condition.

Theorem 6.10 *Let $A \subset \mathbb{R}^2$ be a 2-dimensional piecewise linear manifold with boundary $X = \partial A$. Fix $v \in S^1$. The degree-0 persistent homology of $h_v^X : X \to \mathbb{R}$ can be written as*

$$PH_0(X, h_v^X) = \oplus_{i=1}^m \mathcal{I}_{[h_v(y_{j_i}), d_i)}$$

where $y_{j_1}, \ldots y_{j_m}$ are vertex representatives for each persistent 0-cycle, and $d_1, \ldots d_m \in \mathbb{R} \cup \infty$. Here we have only included intervals with positive length.

Let J^{ord} be the subset of $\{1, 2, \ldots m\}$ such that d_i is finite and y_{j_i} is $(+)$-critical for h_v^A. Then

$$\text{Ord}_0(A, h_v^A) = \oplus_{i \in J^{\text{ord}}} \mathcal{I}_{[(h_v(y_{j_i}), \text{ord}), (d_i, \text{ord}))}.$$

Now let J^{rel} be the subset of $\{1, 2, \ldots m\}$ such that d_i is finite but y_i is not $(+)$-critical for h_v^A.

$$\text{Rel}_1(A, h_v^A) = \oplus_{i \in J^{\text{rel}}} \mathcal{I}_{[(d_i, \text{rel}), (h_v(y_{j_i}), \text{rel}))}.$$

Proof If h_v^A is a Morse function the result follows directly from Theorem 4.18, so suppose that h_v^A is not a Morse function. Recall that since $A \subset \mathbb{R}^2$ is a 2-dimensional piecewise linear manifold with boundary $X = \partial A$, a sufficient condition for h_v^A to be Morse will be that all the vertices in X have distinct values under h_v^A.

Let v_t be the rotation of v anticlockwise by t. Given v there is an $\epsilon > 0$ such that for all $t < \epsilon$ we have $h_v(x) < h_v(y)$ implies $h_{v_t}(x) < h_{v_t}(y)$. We can now break the ties that imply h_v^A is not Morse; where $h_v(x) = h_v(y)$ we have $h_{v_t}(x) \neq h_{v_t}(y)$. We choose $\epsilon > 0$ small enough that $h_{v_t}^A$ is a Morse function for all $t < \epsilon$.

A vertex y_{j_i} will be an isolated 0-critical vertex for h_v^X if and only if it is an isolated vertex for $h_{v_t}^X$, as the order of the heights of y_{j_i-1}, y_{j_i} and y_{j_i+1} are the same under both h_v and h_{v_t}. Since $DET(y_{j_i+1} - y_{j_i}, y_{j_i-1} - y_{j_i}) > 0$ is independent of v we know that whether or not y_{j_i} is $(+)$-critical is the same under h_v^A and $h_{v_t}^A$ by Lemma 6.8. For notational purposes we set $x_{j_i} = y_{j_i}$ to be the vertex representative for $h_{v_t}^X$.

Now suppose $e(x_k, x_l)$ is a 0-critical segment for h_v^X with $y_{j_i} \in e(x_k, x_l)$. Since $h_{v_t}^A$ is Morse, all the vertices in $e(x_k, x_l)$ take distinct values for $h_{v_t}^X$, with exactly one of x_k or x_l now an isolated 0-critical point. Denote this endpoint by x_{j_i}. Since we choose v_t to be a small anticlockwise rotation of v, this choice will be a consistent tie-break for all $0 < t < \epsilon$. Again since $DET(x_{j_i+1} - x_{j_i}, x_{j_i-1} - x_{j_i}) > 0$ is independent of v we know that whether or not x_{j_i} is $(+)$-critical is the same under h_v^A and $h_{v_t}^A$ by Lemma 6.8. Note that by construction, $h_v(x_{j_i}) = h_v(y_{j_i})$.

The above arguments show that we have

$$\text{PH}_0(X, h_v^X) = \oplus_{i=1}^m \mathcal{I}_{[h_v(x_{j_i}), d_i)}.$$

The remainder of the proof is an argument in continuity. For $t \in (0, \epsilon)$ we have

$$\text{PH}_0(X, h_{v_t}^X) = \oplus_{i=1}^m \mathcal{I}_{[h_{v_t}(x_{j_i}), d_{i,t})}$$

for some $d_{i,t} \in \mathbb{R}$. Since $\lim_{t \to 0^+} v_t = v$ we have $\lim_{t \to 0^+} \text{PH}_0(X, h_{v_t}^X) = \text{PH}_0(X, h_v^X)$ and thus $\lim_{t \to 0^+} d_{i,t} = d_i$ for all i.

Since each x_{j_i} is $(+)$-critical with respect to h_v^A if and only if it is $(+)$-critical with respect to $h_{v_t}^A$ we can apply Theorem 4.18 to say for all $t \in (0, \epsilon)$ that

$$\text{Ord}_0(A, h_{v_t}^A) = \oplus_{i \in J^{\text{ord}}} \mathcal{I}_{[(h_{v_t}(y_{j_i}), \text{ord}), (d_{i,t}, \text{ord}))}$$

and

$$\mathrm{Rel}_1(A, h_{v_t}^A) = \oplus_{i \in J^{\mathrm{rel}}} \mathcal{I}_{[(d_{i,t}, \mathrm{rel}), (h_{v_t}(y_{j_i}), \mathrm{rel}))}.$$

Taking the limit as $t \to 0^+$ completes the proof. □

7 Implementation details

Algorithm 1 Computing the XPHT from a binary image.

Require: A binary image $\mathcal{I}$, number of directions $K = 4k$.

1: Extract directed edges from $\mathcal{I}$

2: Form boundary curves from directed edges

3: Assign interior and exterior labels to boundary curves

4: Initialise Ord0, Rel1, Ess0, Ess1 as lists of length $4k$ with each element an empty list.

5: **for** $i = 1, \ldots 2k$ **do**

6:      $v = (\cos(\frac{\pi i}{2k}), \sin(\frac{\pi i}{2k}))$

7:      **for** each boundary curve **do**

8:          compute degree-0 persistence pairs for the lower-level set filtration of $h_v(x) = \langle x, v \rangle$

9:          assign $(+)$-critical or $(-)$-critical flags to each finite-lifetime persistence pair

10:          determine maximum and minimum values and critical points for $h_v(x)$

11:          Ord0$(i) \leftarrow (+)$-critical persistence pairs   ▷ the symbol $\leftarrow$ here means 'append'

12:          Rel1$(i) \leftarrow (-)$-critical persistence pairs

13:          Ord0$(i + 2k) \leftarrow$ reversed $(-)$-critical persistence pairs

14:          Rel1$(i + 2k) \leftarrow$ reversed $(+)$-critical persistence pairs

15:          **if** boundary curve is exterior **then**

16:              Ess0$(i) \leftarrow$ min, max pair

17:              Ess0$(i + 2k) \leftarrow$ -max, -min pair

18:          **if** boundary curve is interior **then**

19:              Ess1$(i) \leftarrow$ max, min pair

20:              Ess1$(i + 2k) \leftarrow$ -min, -max pair

21: **return** Ord0, Rel1, Ess0, Ess1

Fig. 6 The input binary image A with foreground in grey. The boundary curves ∂A are oriented anticlockwise with the interior curve in orange and the exterior curves in blue. These curves are constructed using the rules illustrated in Fig. 5 (color figure online)
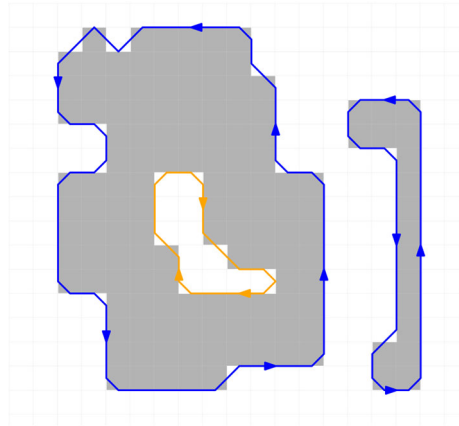

Using the theory developed in the previous sections we have implemented a package in R which takes as input a binary image and outputs the extended persistent homology transform of the foreground of that image. The R-package is available at https://github.com/james-e-morgan/xpht. The paragraphs below describe a simple example to illustrate the sequence of steps followed when using the package. We finish this section with a fun application using the XPHT to cluster the shapes of letters from various standard fonts.

Let A denote the foreground of the binary image and X the boundary between the foreground and background as constructed in the previous section. The user chooses the number of directions K, and the unit vectors are set to $v_i = (\cos(2\pi i/K), \sin(2\pi i)/K)$. We can compute the extended persistent homology of A for directions v and $-v$ from the regular persistent homology of X in direction v together with knowledge of the minimum and maximum values of h_v^X on each boundary curve. Therefore, when the number of directions is even, the computational time for the XPHT is halved. If the user has a collection of shapes that require centring (Turner et al. 2014) then K is required to be a multiple of four. The main steps used in the R-code are summarised in Algorithm 1 and illustrated with an example below.

The first step is to construct the oriented boundary curves around each of the components, labelling which curves are interior and which are exterior. Note that by Lemma 6.3 the boundary between the foreground and background is a collection of disjoint closed curves. This set of boundary curves is independent of direction vector and is computed only once. For an example see Fig. 6. Constructing the set of all oriented edges using the 2×2 image patch lookup table in Fig. 5 requires $O(N)$ steps where N is the number of pixels. Building the boundary curves as ordered lists of vertices from these edges is $O(n)$ where $n \leq 2N$ is the number of edges. Determining the interior/exterior status of a boundary curve is $O(1)$. The overall cost of constructing the oriented boundary curves of the foreground is therefore $O(N)$ where N is the number of pixels.

For each direction v the regular degree-0 persistent homology of the boundary curves can be computed very efficiently using the union-find data structure. The com-

Fig. 7 The critical points for $PH_0(\partial A, v)$ for the given direction v. The vertices marked with crosses are 0-critical points and correspond to birth events in $PH_0(\partial A, v)$. Vertices marked with circles are 1-critical and cause a death in $PH_0(\partial A, v)$. The same letter label is given to the paired birth and death events of a persistent homology class from $PH_0(\partial A, v)$
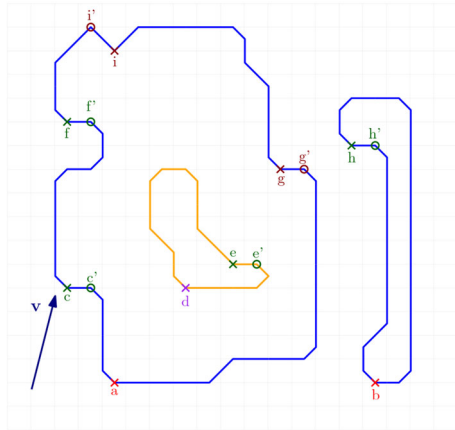
Fig. 8 Identifying the boundary curve local minima that are also local minima of h_v on the foreground. The 0-critical points for X that correspond to births of finite lifetime persistence classes in $PH_0(X, v)$ are c, e, f, g, h and i. We have of these c, e, f and h are local minima for A and thus (+)-critical points. The remaining (g and i) are (−)-critical
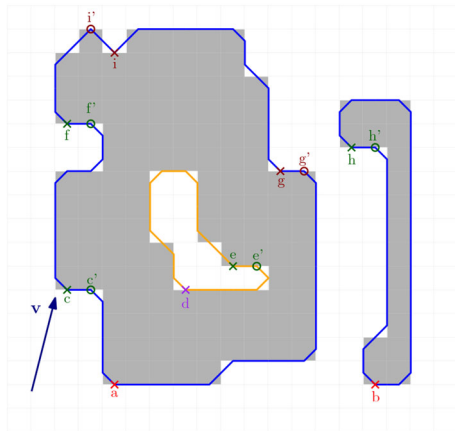
plexity is known to be $O(nA^{-1}(n))$ where $A^{-1}(n)$ is the inverse Ackermann function and n is the total number of boundary edges. Our implementation also identifies the 0-critical vertices that represent births of components of ∂A with respect to the filtration direction v; see Fig. 7.

Using Lemma 6.8 or 6.9 we determine which 0-critical points are positive critical or negative critical for the foreground and label the ordinary persistent homology classes as either (+)-critical or (−)-critical. This is illustrated for the example in Fig. 8. Determining the sign at each critical point has constant complexity, $O(1)$.

Using Theorem 6.10 we can compute the ordinary and relative persistent homology for h_v^A from the persistent homology of h_v^X together with information about which 0-critical isolated vertices and 0-critical segments are (+)-critical. Applying the duality result from Corollary 2.7 we deduce the ordinary and relative persistence modules for direction $-v$ from those for direction v. In our worked example:

$$\mathrm{Ord}_0(A, v) = \mathcal{I}_{[(h_v(c),\mathrm{ord}),(h_v(c'),\mathrm{ord})]} \oplus \mathcal{I}_{[(h_v(e),\mathrm{ord})(h_v(e'),\mathrm{ord})]}$$
$$\oplus \mathcal{I}_{[(h_v(f),\mathrm{ord})(h_v(f'),\mathrm{ord})]} \oplus \mathcal{I}_{[(h_v(h),\mathrm{ord})(h_v(h'),\mathrm{ord})]}$$

Fig. 9 The minima and maxima points of h_v^X for each boundary curve in X. The exterior curves have (minimum,maximum) pairs labelled (m_1, M_1) and (m_2, M_2) and the interior curve has the pair of (m_3, M_3)
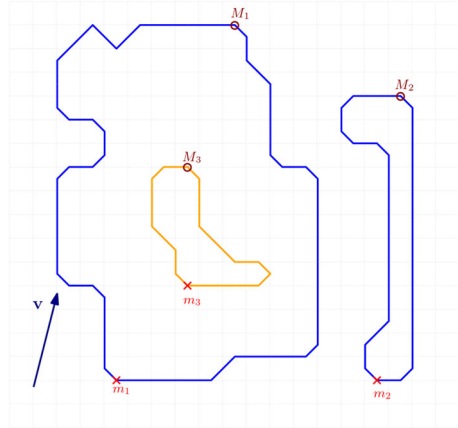
$$\text{Rel}_1(A, v) = \mathcal{I}_{[(h_v(i'),\text{rel}),(h_v(i),\text{rel})]} \oplus \mathcal{I}_{[(h_v(g'),\text{rel})(h_v(g),\text{rel}))}$$
$$\text{Ord}_0(A, (-v)) = \mathcal{I}_{[(-h_v(i'),\text{ord}),(-h_v(i),\text{ord}))} \oplus \mathcal{I}_{[(-h_v(g'),\text{ord})(-h_v(g),\text{ord}))}$$
$$\text{Rel}_1(A, (-v)) = \mathcal{I}_{[(-h_v(c),\text{rel}),(-h_v(c'),\text{rel}))} \oplus \mathcal{I}_{[(-h_v(e),\text{rel})(-h_v(e'),\text{rel}))}$$
$$\oplus \mathcal{I}_{[(-h_v(f),\text{rel})(-h_v(f'),\text{rel}))} \oplus \mathcal{I}_{[(-h_v(h),\text{rel})(-h_v(h'),\text{rel}))}.$$

To compute the essential classes we use Proposition 4.20. Each of the boundary curves is labelled as interior or exterior. We compute the essential classes by finding the minimum and maximum values of h_v^X on these boundary curves. This is illustrated for our running example in Fig. 9.

Using the notation of the figure, the essential classes for the foreground A are therefore

$$\text{Ess}_0(A, v) = \mathcal{I}_{[(h_v(m_1),\text{ord}),(h_v(M_1),\text{rel})} \oplus \mathcal{I}_{[(h_v(m_2),\text{ord}),(h_v(M_2),\text{rel})}$$

and

$$\text{Ess}_1(A, v) = \mathcal{I}_{[(h_v(M_3),\text{ord}),(h_v(m_3),\text{rel})}.$$

We then infer the essential persistence modules for direction $-v$ to be

$$\text{Ess}_0(A, -v) = \mathcal{I}_{[(-h_v(M_1),\text{ord}),(-h_v(m_1),\text{rel})} \oplus \mathcal{I}_{[(-h_v(M_2),\text{ord}),(-h_v(m_2),\text{rel})}$$

and

$$\text{Ess}_1(A, -v) = \mathcal{I}_{[(-h_v(m_3),\text{ord}),(-h_v(M_3),\text{rel})}.$$

The overall complexity of the algorithm is $O(N)$ for computing the oriented boundary curves, where N is the number of pixels, and then $2k\,O(n)$ to compute the extended persistent homology in $4k$ directions, where n is the total number of edges in the boundary curves.
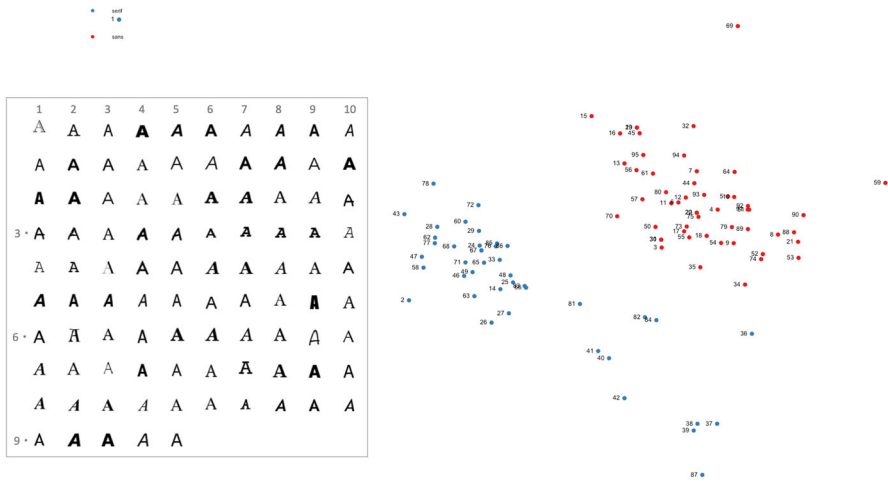
Fig. 10 Upper case 'A' rendered in a variety of fonts. The letter shapes are numbered 1–95 reading left to right, top to bottom in the 10 by 10 grid. The 1-Wasserstein distances between each pair of letters are visualised using MDS with two dimensions. This shows a separation between serif 'A' and sans serif 'A' fonts plotted with blue and red respectively. The upper outlier labelled 69 is 'Noteworthy-Light', a large round simple script font. The lower outlier labelled 87 is 'Trattatello' and has the smallest height and counter in this set. In the upper left corner hiding in the legend is letter 1, 'Academy-Engraved' which has outlined strokes giving it three holes. On the far right is letter 59 ('Impact'), notable for having a narrow body width but heavy weight (color figure online)

Example 7.1 We now briefly describe results from an XPHT analysis of the capital letter A and the lower-case letter g rendered using over 90 standard fonts. Each letter was created as a small binary image (130×130 pixels) using an 84pt font size; these are shown in Figs. 10 and 11. The XPHT for each letter was computed using $K = 32$ directions. Fonts vary in their letter placement with respect to a baseline, so we centred the XPHT summary for each shape using the method outlined in Turner et al. (2014). We did not need to consider angular alignment as the images are generated with a consistent orientation. The letters all have the same specified font size so we discuss the XPHT results without any scaling which allows the different heights and widths to serve as characteristics of the font. We can also scale the XPHT summaries so that each letter has the same diameter which changes the pairwise distances between these shapes a little; the results are given as supplementary figures in the appendix.

We computed all pairwise distances between the XPHT summaries using both the 1- and 2-Wasserstein metrics. Results for the 1-Wasserstein case are presented and discussed here; the other plots are given in the appendix. To demonstrate the types of shape features the XPHT captures, we use multi-dimensional scaling (MDS) to assign planar coordinates to each letter. The plots in Figs. 10 and 11 show that the XPHT distances capture the difference between serif and sans-serif versions of the letter A, and between single- and double-storey versions of the letter g. Of particular note is the font 'Chalkduster' (label 32) which has a textured look with small holes and rough boundary; the XPHT distances don't make this a significant outlier for the letter As. Chalkduster g is an outlier for that set because the bowl doesn't create a

Fig. 11 Lower case 'g' rendered in the same fonts and same order as used for 'A'. The 1-Wasserstein distances between each pair of letters are again visualised using MDS in two dimensions. In this case, there is good separation between single storey 'g' and double storey 'g' font shapes. The upper right outlier labelled 32 is 'Chalkduster' and has no essential 1-cycles. The lower left outlier labelled 1 is 'Academy Engraved', it has additional essential 1-cycles due to its outlined stroke style. There is one green data point, 36 'Copperplate', which is rendered in upper-case form. The red, double-storey 'g' fonts form two distinct clusters. The fonts placed to the right are those where the lower tail doesn't quite form a closed loop. At the lower right are letters 15 and 16 (forms of 'Avantgarde'); these have the roundest bowls and largest counters, i.e., large circular upper holes (color figure online)

closed 1-cycle. It's also worth noting that the XPHT distances create two clusters for the double-storey letter 'g's with those that have $\beta_1 = 2$ on the left and those, such as fonts labelled 62 and 89, which look double-storey but have $\beta_1 = 1$ clustered on the right and closer to the single-storey fonts. We emphasize that an ordinary PHT shape analysis could not replicate this clustering because the distances between single-storey and true double-storey 'g's would be infinite.

These letters are included in the R-package release and more details about the analysis are provided in the vignettes.

8 Future directions

This paper presents a new approach to computing persistent homology for manifolds with boundary by exploiting relationships between the extended persistent homology of a manifold with boundary to that of just the boundary. Although the focus here has been on height functions of embedded shapes in Euclidean space it is reasonable to expect that similar results could hold for other kinds of functions, such as radial functions. One application already described, is the use of the XPHT in detecting symmetries of 2D shapes (Bermingham et al. 2023). Future directions of research also include considering generalisations to stratified spaces, adapting ideas from stratified Morse theory as developed by Goresky and MacPherson (1988).

Other areas to explore are theoretical properties of the XPHT. Stability results for the PHT for different embeddings of the same simplicial complex were proved in Skraba and Turner (2020) and these should hold for the XPHT by the same arguments. More generally, we expect better stability results for the XPHT than for the PHT as we can introduce new essential classes with small support without dramatically changing the extended persistent homology transform.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no Conflict of interest.

References

Amézquita, E.J., Quigley, M.Y., Ophelders, T., Landis, J.B., Koenig, D., Munch, E., Chitwood, D.H.: Measuring hidden phenotype: quantifying the shape of barley seeds using the Euler characteristic transform. In Silico Plants **4**(1), diab033 (2022)

Bauer, U., Lesnick, M.: Induced matchings of barcodes and the algebraic stability of persistence. In Proceedings of the Thirtieth Annual Symposium on Computational Geometry, pp. 355–364 (2014)

Bauer, U., Botnan, M.B., Fluhr, B.: Universality of the bottleneck distance for extended persistence diagrams (2020). arXiv:2007.01834

Bermingham, Nicholas, Robins, Vanessa, Turner, Katharine: Planar symmetry detection and quantification using the extended persistent homology transform. In *Proceedings of IEEE Vis 2023 TopoInVis workshop*, pages 1–9, Melbourne, Australia, (2023). IEEE

Braess, D.: Morse-theorie für berandete mannigfaltigkeiten. Math. Ann. **208**(2), 133–148 (1974)

Carlsson, G., De Silva, V., Kališnik, S., Morozov, D.: Parametrized homology via zigzag persistence. Algebraic Geom. Topol. **19**(2), 657–700 (2019)

Chazal, F., Crawley-Boevey, W., De Silva, V.: The observable structure of persistence modules (2014). arXiv:1405.5644

Cohen-Steiner, D., Edelsbrunner, H., Harer, J.: Erratum to: Extending persistence using Poincaré and Lefschetz duality. Found. Comput. Math. **9**(1), 133–134 (2009)

Cohen-Steiner, D., Edelsbrunner, H., Harer, J.: Extending persistence using Poincaré and Lefschetz duality. Found. Comput. Math. **9**(1), 79–103 (2009)

Crawford, L., Monod, A., Chen, A.X., Mukherjee, S., Rabadán, R.: Predicting clinical outcomes in glioblastoma: an application of topological and functional data analysis. J. Am. Stat. Assoc. **115**(531), 1139–1150 (2020)

Crawley-Boevey, W.: Decomposition of pointwise finite-dimensional persistence modules. J. Algebra Appl. **14**(05), 1550066 (2015)

Curry, J., Mukherjee, S., Turner, K.: How many directions determine a shape and other sufficiency results for two topological transforms. Trans. Am. Math. Soc. Ser. B **9**(32), 1006–1043 (2022)

De Silva, V., Morozov, D., Vejdemo-Johansson, M.: Dualities in persistent (co) homology. Inverse Prob. **27**(12), 124003 (2011)

Edelsbrunner, H., Mücke, E.P.: Simulation of simplicity: a technique to cope with degenerate cases in geometric algorithms. ACM Trans. Graph. **9**(1), 66–104 (1990)

Edelsbrunner, H., Kerber, M.: Alexander duality for functions: the persistent behavior of land and water and shore. In Proceedings of the Twenty-Eighth Annual Symposium on Computational Geometry, SoCG '12, pp. 249–258, New York, NY, USA (2012). ACM

Ghrist, R., Levanger, R., Mai, H.: Persistent homology and Euler integral transforms. J. Appl. Comput. Topol. **2**(1), 55–60 (2018)

Goresky, M., MacPherson, R.: Stratified Morse theory. In Stratified Morse Theory, pp. 3–22. Springer (1988)

Grunert, R., Kühnel, W., Rote, G.: Pl Morse theory in low dimensions (2019). arXiv:1912.05054

Jankowski, A, Rubinsztejn, E: Functions with non-degenerate critical points on manifolds with boundary. Commentationes Mathematicae, **16**(1), (1972)

Kong, T.Y., Rosenfeld, A.: Digital topology: introduction and survey. Comput. Vis. Graph. Image Process. **48**(3), 357–393 (1989)

Milnor, J.: Morse theory. Based on lecture notes by M. Spivak and R. Wells. Annals of Mathematics Studies, No. 51. Princeton University Press, Princeton, NJ (1963)

Newman, T.S., Yi, H.: A survey of the marching cubes algorithm. Comput. Graph. **30**(5), 854–879 (2006)

Shboul, Z.A., Alam, M., Vidyaratne, L., Pei, L., Elbakary, M.I., Iftekharuddin, K.M.: Feature-guided deep radiomics for glioblastoma patient survival prediction. Front. Neurosci. 966 (2019)

Skraba, P., Turner, K.: Wasserstein stability for persistence diagrams (2020). arXiv:2006.16824

Tang, W.S., da Silva, G.M., Kirveslahti, H., Skeens, E., Feng, B., Sudijono, T., Yang, K.K., Mukherjee, S., Rubenstein, B., Crawford, L.: A topological data analytic approach for discovering biophysical signatures in protein dynamics. PLoS Comput. Biol. **18**(5), e1010045 (2022)

Turner, K., Mukherjee, S., Boyer, D.M.: Persistent homology transform for modeling shapes and surfaces. Inform. Inference: J. IMA **3**(4), 310–344 (2014)

Zhang, Y., Peng, J., Yuan, X., Zhang, L., Zhu, D., Hong, P., Wang, J., Liu, Q., Liu, W.: Mfcis: an automatic leaf-based identification pipeline for plant cultivars using deep learning and persistent homology. Horticult. Res. **8**, (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.